Chapter 10
Preliminaries for Nonparametric Regression

10.1 Introduction

In all other chapters we assume that the regression model, $f(\boldsymbol{x})$, takes an a priori specified, usually simple, parametric form. Such models have a number of advantages: If the assumed parametric form is approximately correct, then efficient estimation will result; having a specific linear or nonlinear form allows concise summarization of an association; inference for parametric models is often relatively straightforward. Further, a particular model may be justifiable from the context.

In this and the following two chapters, we consider situations in which a greater degree of flexibility is desired, at least when modeling some components of the covariate vector $\boldsymbol{x}$. Nonparametric modeling is particularly useful when one has little previous experience with the specific data-generating context. Typically, one may desire $f(\cdot)$ to arise from a class of functions with restrictions on smoothness and continuity. Although the models of this and the next two chapters are referred to as nonparametric,[1] they often assume parametric forms but depend on a large number of parameters which are constrained in some way, in order to prevent overfitting of the data. For some approaches, for example, the regression tree models described in Sect. 12.7, the model is specified implicitly through an algorithm, with the specific form (including the number of parameters) being selected adaptively.

There are a number of contexts in which flexible modeling is required. The simplest is when a graphical description of a set of data is needed, which is often referred to as *scatterplot smoothing*. Formal inference is also possible within a nonparametric framework, however. In some circumstances, estimation of a parametric relationship between a response and an x variable may be of interest, while requiring flexible nonparametric modeling of other nuisance variables (including confounders). The example described in Sect. 1.3.6 is of this form, with the association between spinal bone mineral density and ethnicity being of primary

[1] Some authors prefer the label *semiparametric*.

J. Wakefield, *Bayesian and Frequentist Regression Methods*, Springer Series in Statistics, DOI 10.1007/978-1-4419-0925-1_10,

© Springer Science+Business Media New York 2013

interest, but with a flexible model for age being desired. Finally, an important and common use of nonparametric modeling is prediction. In this case, the focus is on the accuracy of the final prediction, with little interest in the values of the parameters in the model. Prediction with a discrete outcome is often referred to as *classification*.

Much of the development of nonparametric methods, in particular those associated with classification, has occurred in computer science and, more specifically, machine learning, with a terminology that is quite different to that encountered in the statistics literature. The data with which the model is fitted constitute the *training sample*; nonparametric regression is referred to as *learning a function*; the covariates are called *features*; and adding a penalty term to an objective function (e.g., a residual sum of squares) is called *regularization*. In *supervised learning* problems, there is an outcome variable that we typically wish to predict, while in *unsupervised learning* there is no single outcome to predict, rather the aim is to explore how the data are organized or clustered. Only supervised learning is considered here.

The layout of this chapter is as follows. In Sect. 10.2, we discuss a number of motivating examples. Section 10.3 examines what response summary should be reported in a prediction setting using a decision theory framework, while in Sect. 10.4 various measures of predictive accuracy are reviewed. A recurring theme will be the bias-variance trade-off encountered when fitting flexible models containing a large number of parameters. To avoid excess variance of the prediction, various techniques that reduce model complexity will be described; a popular approach is to penalize large values of the parameters. This concept is illustrated in Sect. 10.5 with descriptions of *ridge regression* and the *lasso*. These *shrinkage* methods are introduced in the context of multiple linear regression.[2] Controlling the complexity of a model is a key element of nonparametric regression and is usually carried out using smoothing (or tuning) parameters. In Sect. 10.6, smoothing parameter estimation is considered. Concluding comments appear in Sect. 10.7. There is a huge and rapidly growing literature on nonparametric modeling, and the surface is only scratched here; Sect. 10.8 gives references to broader treatments and to more detailed accounts of specific techniques.

The next two chapters also consider nonparametric modeling. In Chap. 11, two popular approaches to smoothing are described: Those based on splines and those based on kernels; the focus of the latter is local regression. Chapter 11 only considers situations with a single covariate, with multiple predictors considered in Chap. 12, along with methods for classification.

10.2 Motivating Examples

Three examples that have been previously introduced will be used for illustrating nonparametric modeling: The prostate cancer data described in Sect. 1.3.1 are used for illustration in this chapter and in Chap. 12; the spinal bone marrow data of

[2] Ridge regression is also briefly encountered in Sect. 5.12.

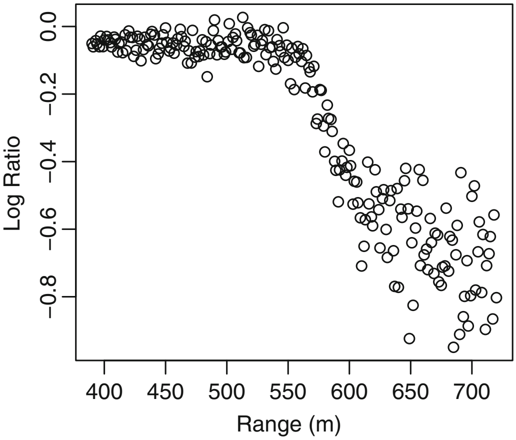


Fig. 10.1 Log ratio of two laser sources, as a function of the range, in the LIDAR data

Sect. 1.3.6 will be analyzed in Chap. 11; and the bronchopulmonary dysplasia data described in Sect. 7.2.3 will be examined in Chaps. 11 and 12. In this section, two additional datasets are described.

10.2.1 Light Detection and Ranging

Figure 10.1 shows data, taken from Holst et al. (1996), from a light detection and ranging (LIDAR) experiment. The LIDAR technique (which is similar to radar technology) uses the reflection of laser-emitted light to monitor the distribution of atmospheric pollutants. The data we consider concern mercury. The x-axis measures distance traveled before light is reflected back to its source (and is referred to as the range), and the y-axis is the logarithm of the ratio of distance measured for two laser sources: One source has a frequency equal to the resonant frequency of mercury, and the other has a frequency off this resonant frequency. For these data, point and interval estimates for the association between the log ratio and range are of interest. Figure 10.1 shows a clear nonlinear relationship between the log ratio and range, with greater variability at larger ranges.

10.2.2 Ethanol Data

This example concerns data collected in a study reported by Brinkman (1981). The data consist of $n = 88$ measurements on three variables: NOx, the concentration of nitric oxide (NO) and nitrogen dioxide (NO_2) in the engine exhaust, with normalization by the work done by the engine; C, the compression ratio of the engine; and E, the equivalence ratio at which the engine was run, a measure of

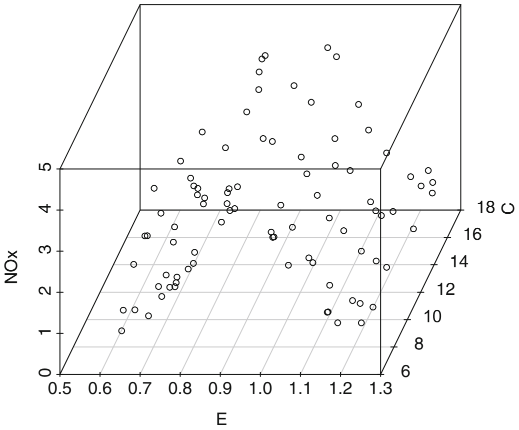


Fig. 10.2 A three-dimensional display of the ethanol data, showing the normalized concentration of nitric oxide and nitrogen dioxide (NOx) as a function of the equivalance ratio at which the engine was run (E) and the compression ratio of the engine (C)

the air/ethanol mix. Figure 10.2 gives a three-dimensional display of these data. The aim is to build a predictive model, and a simple linear model is clearly inadequate since there is a strong nonlinear (inverse U-shaped) association between NOx and E. The form of the association between NOx and C is less clear.

10.3 The Optimal Prediction

Before considering model specification and describing methods for fitting, we use a decision theory framework to decide on which summary of the distribution of $Y \mid \boldsymbol{x}$ we should report if the aim of analysis is prediction, where $\boldsymbol{x}$ is a $1 \times (k+1)$ design vector corresponding to the intercept and k covariates. Throughout this section, we will suppose we are in an idealized situation in which all aspects of the data-generating mechanism are known, and we need only decide on which quantity to report.

The specific decision problem we consider is the following. Imagine we are involved in a game in which the aim is to predict a new observation y, using a function of covariates $\boldsymbol{x}$, $f(\boldsymbol{x})$. Further, we know that our predictions will be penalized via a loss function $L[y, f(\boldsymbol{x})]$ that is the penalty incurred when predicting y by $f(\boldsymbol{x})$. The optimal prediction is that which minimizes the expected loss defined as

$$\mathrm{E}_{\boldsymbol{X},Y}\left\{L\left[Y, f(\boldsymbol{X})\right]\right\}, \tag{10.1}$$

where the expectation is with respect to the joint distribution of the random variables Y and $\boldsymbol{X}$.

10.3.1 Continuous Responses

The most common choice of loss function is squared error loss, with $f(\boldsymbol{x})$ chosen to minimize the expected (squared) prediction error:

$$\mathrm{E}_{\boldsymbol{X},\,Y}\left\{[Y - f(\boldsymbol{X})]^2\right\}, \tag{10.2}$$

that is, the quadratic loss. Writing (10.2) as

$$\mathrm{E}_{\boldsymbol{X}}\left[\mathrm{E}_{Y\,|\,\boldsymbol{X}\,=\,\boldsymbol{x}}\left\{[Y - f(\boldsymbol{x})]^2 \mid \boldsymbol{X} = \boldsymbol{x}\right\}\right]$$

indicates that we may minimize pointwise, with solution

$$\widehat{f}(\boldsymbol{x}) = \mathrm{E}[Y \mid \boldsymbol{x}],$$

that is, the conditional expectation (Exercise 10.4). Hence, the best prediction, $\widehat{f}(\boldsymbol{x})$, is the usual regression function.

As an alternative, with absolute loss, $\mathrm{E}_{\boldsymbol{X},\,Y}[\ |Y - f(\boldsymbol{X})|\ ]$, the solution is the conditional median

$$\widehat{f}(\boldsymbol{x}) = \mathrm{median}(Y \mid \boldsymbol{x})$$

(Exercise 10.4). Modeling via the median, rather than the mean, provides greater robustness to outliers but with an increase in computational complexity. Exercise 10.4 also considers a generalization of absolute loss.

Other choices have also been suggested for specific situations. For example, the scaled quadratic loss function

$$L[y, f(\boldsymbol{x})] = \left(\frac{y - f(\boldsymbol{x})}{y}\right)^2 \tag{10.3}$$

has been advocated for random variables $y > 0$ (e.g., Bernardo and Smith 1994, p. 301). This loss function is scaling departures $y - f(\boldsymbol{x})$ by y, so that discrepancies in the predictions of the same magnitude are penalized more heavily for small y than for large y. Taking the expectation of (10.3) with respect to $Y \mid \boldsymbol{x}$ leads to

$$\widehat{f}(\boldsymbol{x}) = \frac{\mathrm{E}[Y^{-1} \mid \boldsymbol{x}]}{\mathrm{E}[Y^{-2} \mid \boldsymbol{x}]}. \tag{10.4}$$

For details, see Exercise 10.5. As an example, suppose the data are gamma distributed as

$$Y \mid \mu(\boldsymbol{x}), \alpha \sim_{iid} \mathrm{Ga}\left\{\alpha^{-1}, [\mu(\boldsymbol{x})\alpha]^{-1}\right\},$$

where $\mathrm{E}[Y \mid \boldsymbol{x}] = \mu(\boldsymbol{x})$, and $\alpha^{-1/2}$ is the coefficient of variation. Then Exercise 10.5 shows that (10.4) is equal to

$$\widehat{f}(\boldsymbol{x}) = (1 - 2\alpha)\mu(\boldsymbol{x}), \tag{10.5}$$

for $\alpha < 0.5$. Hence, under the scaled quadratic loss function, we should scale the mean function by $1 - 2\alpha$ when reporting.

10.3.2 Discrete Responses with K Categories

Now suppose the response is categorical, with $Y \in \{0, 1, \ldots, K-1\}$. Again, we must decide on which summary measure to report. One approach is to assign a class in $\{0, 1, \ldots, K-1\}$ to a new case via a classification rule $g(\boldsymbol{x})$. Alternatively, a probability distribution over the classes may be reported.[3]

Suppose the distributions of $\boldsymbol{x}$ given $Y = k$, $p(\boldsymbol{x} \mid Y = k)$, are known along with prior probabilities on the classes, $\Pr(Y = k) = \pi_k$. Then, via Bayes theorem, the posterior classifications may be obtained:

$$\Pr(Y = k \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid Y = k)\pi_k}{\sum_{l=0}^{K-1} p(\boldsymbol{x} \mid Y = l)\pi_l}. \tag{10.6}$$

Choosing the k that maximizes these probabilities gives a *Bayes classifier*.

For the situation in which we wish to assign a class label, the loss function is a $K \times K$ matrix $\boldsymbol{L}$ with element $L(j,k)$ representing the loss incurred when the truth is $Y = j$, and the classification is $g(\boldsymbol{x}) = k$, with $j, k \in \{0, 1, \ldots, K-1\}$. A sensible loss function is

$$L(j,k) = \begin{cases} 0 & \text{if } j = k \\ \geq 0 & \text{if } j \neq k. \end{cases} \tag{10.7}$$

In most cases,we will assign $L(j,k) > 0$ for $j \neq k$ but in some contexts incorrect classifications will not be penalized if they are of no consequence. We emphasize that the *class* predictor $g(\boldsymbol{x})$ takes a value from the set $\{0, 1, \ldots, K-1\}$ and is a function of $\Pr(Y = k \mid \boldsymbol{x})$. The expected loss is

$$\begin{aligned} \mathrm{E}_{\boldsymbol{X},Y}\left\{L\left[Y, g(\boldsymbol{X})\right]\right\} &= \mathrm{E}_{\boldsymbol{X}}\left[\mathrm{E}_{Y|\boldsymbol{x}}\left\{L\left[Y, g(\boldsymbol{x})\right] \mid \boldsymbol{X} = \boldsymbol{x}\right\}\right] \\ &= \mathrm{E}_{\boldsymbol{X}}\left[\sum_{k=0}^{K-1} L\left[Y = k, g(\boldsymbol{x})\right]\Pr(Y = k \mid \boldsymbol{x})\right]. \end{aligned} \tag{10.8}$$

[3] It is possible to also have a “doubt” category that is assigned if there is sufficient ambiguity but we do not consider this possibility. See Ripley (1996) for further discussion.

Table 10.1 Loss table for a binary decision problem

		Predicted Class	
		$g(\boldsymbol{x}) = 0$	$g(\boldsymbol{x}) = 1$
True	$Y = 0$	0	$L(0, 1)$
Class	$Y = 1$	$L(1, 0)$	0

where we are assuming the form of $g(\boldsymbol{x})$ is known. The inner expectation of (10.8) is known as the *Bayes risk* (e.g., Ripley 1996), with minimum

$$\widehat{g}(\boldsymbol{x}) = \text{argmin}_{g(x)\in\{0,\dots,K-1\}} \sum_{k=0}^{K-1} L\left[Y = k, g(\boldsymbol{x})\right] \Pr(Y = k \mid \boldsymbol{x}).$$

The $K = 2$ situation will now be considered in greater detail. Table 10.1 gives the table of losses for this case. The Bayes risk is minimized by the choice

$$\widehat{g}(\boldsymbol{x}) =$$
$$\text{argmin}_{g(x)\in\{0,1\}}\{L\left[Y = 0, g(\boldsymbol{x})\right] \Pr(Y = 0 \mid \boldsymbol{x}) + L\left[Y = 1, g(\boldsymbol{x})\right] \Pr(Y = 1 \mid \boldsymbol{x})\} .$$

Hence,

$$\begin{aligned} g(\boldsymbol{x}) = 0 \text{ gives Bayes risk } &= L(1, 0) \times \Pr(Y = 1 \mid \boldsymbol{x}) \\ g(\boldsymbol{x}) = 1 \text{ gives Bayes risk } &= L(0, 1) \times \left[1 - \Pr(Y = 1 \mid \boldsymbol{x})\right] \end{aligned}$$

and so the Bayes risk is minimized by $g(\boldsymbol{x}) = 1$ if

$$L(1, 0) \times \Pr(Y = 1 \mid \boldsymbol{x}) > L(0, 1) \times \left[1 - \Pr(Y = 1 \mid \boldsymbol{x})\right]$$

or equivalently if

$$\frac{\Pr(Y = 1 \mid \boldsymbol{x})}{1 - \Pr(Y = 1 \mid \boldsymbol{x})} > \frac{L(0, 1)}{L(1, 0)} = R \tag{10.9}$$

with the consequence that only the ratio of losses R requires specification. A final restatement is to classify a new case with covariates $\boldsymbol{x}$ as $g(\boldsymbol{x}) = 1$ if

$$\Pr(Y = 1 \mid \boldsymbol{x}) > \frac{L(0, 1)}{L(0, 1) + L(1, 0)} = \frac{R}{1 + R}.$$

If classifying as $g(\boldsymbol{x}) = 1$ when $Y = 0$ is much worse than classifying as $g(\boldsymbol{x}) = 0$ when $Y = 1$, then R should be given a value greater than 1. In this case, if $\Pr(Y = 1 \mid \boldsymbol{x}) > 0.5$ then we assign $g(\boldsymbol{x}) = 1$. For example, if $R = 4$, we set $g(\boldsymbol{x}) = 1$ only if $\Pr(Y = 1 \mid \boldsymbol{x}) > 0.8$.

Returning to the case of general K, in the most straightforward case of all errors being equal, we simply assign an observation to the most likely class, using the probabilities $\Pr(Y = k \mid \boldsymbol{x})$, $k = 0, 1, \dots, K - 1$.

We now turn to the second situation in which a classification is not required, but rather a set of probabilities over $\{0, 1, \ldots, K-1\}$, that is, we require $\boldsymbol{f}(\boldsymbol{x}) = [f_0(\boldsymbol{x}), \ldots, f_{K-1}(\boldsymbol{x})]$. First, consider the $K = 2$ (binary) case. In this case we simplify notation and write $\boldsymbol{f}(\boldsymbol{x}) = [\, f(\boldsymbol{x}), 1 - f(\boldsymbol{x})\,]$. We may specify a loss function which is proportional to the negative Bernoulli log-likelihood

$$L[y, f(\boldsymbol{x})] = -2y \log [f(\boldsymbol{x})] - 2(1-y) \log [1 - f(\boldsymbol{x})] \tag{10.10}$$

where $f(\boldsymbol{x})$ is the function that we will report. Therefore, if the log-likelihood is high the loss is low. The expectation of (10.10) is

$$-2 \Pr(Y = 1 \mid \boldsymbol{x}) \log [f(\boldsymbol{x})] - 2 [1 - \Pr(Y = 1 \mid \boldsymbol{x})] \log [1 - f(\boldsymbol{x})]$$

where $\mathrm{E}[Y \mid \boldsymbol{x}] = \Pr(Y = 1 \mid \boldsymbol{x})$ are the true probabilities, given covariates $\boldsymbol{x}$. The solution is $\widehat{f}(\boldsymbol{x}) = \Pr(Y = 1 \mid \boldsymbol{x})$. Hence, to minimize the expected deviance-type loss function, the true probabilities should be reported, which is not a great surprise.

In the general case of K classes and a multinomial likelihood with one trial and probabilities $\boldsymbol{f}(\boldsymbol{x}) = [f_0(\boldsymbol{x}), \ldots, f_{K-1}(\boldsymbol{x})]$, the deviance loss function is

$$L\,[y, \boldsymbol{f}(\boldsymbol{x})] = -2 \sum_{k=0}^{K-1} I(Y = k) \log f_k(\boldsymbol{x}), \tag{10.11}$$

where $I(\cdot)$ is the indicator function that equals 1 if its argument is true and 0 otherwise. The expected loss is

$$-2 \sum_{k=0}^{K-1} \Pr(Y = k \mid \boldsymbol{x}) \log f_k(\boldsymbol{x})$$

which is minimized by $\widehat{f}_k(\boldsymbol{x}) = \Pr(Y = k \mid \boldsymbol{x})$.

10.3.3 General Responses

In general, if we are willing to speculate on a distribution for the data, we may take the loss function as

$$L[y, f(\boldsymbol{x})] = -2 \log p_f(y \mid \boldsymbol{x}), \tag{10.12}$$

which is the deviance (Sect. 6.5.3), up to an additive constant not depending on f. The notation p_f emphasizes that the distribution of the data depends on f. The previous section gave examples of this loss function for binomial, (10.10), and multinomial, (10.11), data. The loss function (10.12) is an obvious measure of the closeness of y to the predictor function $f(\boldsymbol{x})$ since it is a general measure of the *discrepancy* between the data y and $f(\boldsymbol{x})$. When $Y \mid \boldsymbol{x} \sim \mathrm{N}[f(\boldsymbol{x}), \sigma^2]$, we obtain

$$L[y, f(\boldsymbol{x})] = \log\left(2\pi\sigma^2\right) + [y - f(\boldsymbol{x})]^2 / \sigma^2,$$

which produces $\widehat{f}(\boldsymbol{x}) = \text{E}[Y \mid \boldsymbol{x}]$, as with quadratic loss. Similarly, choosing a Laplacian distribution, that is, $Y \mid \boldsymbol{x} \sim \text{Lap}[f(\boldsymbol{x}), \phi]$ (Appendix D) leads to the posterior median as the optimal choice.

10.3.4 In Practice

Sections 10.3.1 and 10.3.2 describe which summary should be reported, if one is willing to specify a loss function. Such a loss function will often have been based on an implicit model for the distribution of the data or upon an estimation method.

For example, a quadratic loss function is consistent with a model for *continuous responses with additive errors* which is of the form

$$Y = f(\boldsymbol{x}) + \epsilon \tag{10.13}$$

with $\text{E}[\epsilon] = 0$, $\text{var}(\epsilon) = \sigma^2$ and errors on different responses being uncorrelated. This form may be supplemented with the assumption of normal errors or one may simply proceed with least squares estimation. Modeling proceeds by assuming some particular form for $f(\boldsymbol{x})$. A simple approach is to assume that the conditional mean, $f(\boldsymbol{x})$, is approximated by the linear model $\boldsymbol{x\beta}$, as in Chap. 5. Alternative nonlinear models are described in Chap. 6.

Relaxing the constant variance assumption, one may consider generalized linear model (GLM) type situations, to allow for more flexible mean-variance modeling. GLMs are also described in Chap. 6. An assumption of a particular distributional form may be combined with the deviance-type loss function (10.12).

In Sect. 10.3.2 discrete responses were considered, and we saw that with equal losses, one may classify on the basis of the probabilities $\Pr(Y = k \mid \boldsymbol{x})$. As described in Chap. 12, there are two broad approaches to classification. The first approach directly models the probabilities $\Pr(Y = k \mid \boldsymbol{x})$. For example, in the case of binary ($K = 2$) responses, logistic modeling provides an obvious approach (as described in Sect. 7.6). Chap. 12 describes a number of additional methods to model the probabilities as a function of $\boldsymbol{x}$. The second approach is to assume forms for the distributions of $\boldsymbol{x}$ given $Y = k$, $p(\boldsymbol{x} \mid Y = k)$ and then combine these with prior probabilities on the classes, $\Pr(Y = k) = \pi_k$, to form posterior classifications, via (10.6); Chapter 12 also considers this situation.

10.4 Measures of Predictive Accuracy

As already noted, nonparametric modeling is often used for *prediction*, and so the conventional criteria by which methods of parameter estimation are compared (as discussed in Sect. 2.2) are not directly relevant. In a prediction context, there is less concern about the values of the constituent parts of the prediction equation, rather interest is on the total contribution. In Sect. 10.3, loss functions were introduced in order to determine how to report the prediction. In this section, loss functions are used to provide an overall measure of the "error" of a procedure.

The *generalization error* is defined as

$$\text{GE}(\widehat{f}) = \text{E}_{\boldsymbol{X},Y}\left\{L\left[Y, \widehat{f}(\boldsymbol{X})\right]\right\}, \tag{10.14}$$

where $\widehat{f}(\boldsymbol{X})$ is the prediction for Y at a point $\boldsymbol{X}$, with $\boldsymbol{X}, Y$ drawn from their joint distribution. Hence, we are in the so-called X*-random*, as opposed to X*-fixed*, case (Breiman and Spector 1992). The terminology with respect to different measures of accuracy can be confusing and is also inconsistent in the literature; the notation used here is summarized in Table 10.2.

Hastie et al. (2009, Sect. 7.2) describe how one would ideally split the data into three portions with one part being used to fit (or train) models, a second (validation) part to choose a model (which includes both choosing between different classes of models and selecting smoothing parameters within model classes), and a third part to estimate the generalization error of the final model on a test dataset. Unfortunately, there are often insufficient data for division into three parts. Consequently, when prediction methods are to be compared, a common approach is to separate the data into *training* and *test* datasets. The training data are used to train the model and then approximate the validation step using methods to be described in Sect. 10.6. The *test* data are used to estimate the generalization error (10.14) using the function $\widehat{f}(\boldsymbol{x})$ estimated from the training data. We now discuss the form of the generalization error for different data types.

10.4.1 Continuous Responses

To gain flexibility and so minimize bias, predictive models $f(\boldsymbol{x})$ that contain many parameters are appealing. However, if the parameters are not constrained in some way, such models produce wide predictive intervals because a set of data only contains a limited amount of information. In general, as the number of parameters increases, the uncertainty in the estimation of each increases in tandem, which results in greater uncertainty in the prediction also. Consequently, throughout this and the next two chapters, we will repeatedly encounter the bias-variance trade-off. Section 5.9 provides a discussion of this trade-off in the linear model context.

The *expected squared prediction error* is a special case of the generalization error with squared error loss:

$$\text{ESPE}(\widehat{f}) = \text{E}_{\boldsymbol{X},Y}\left\{\left[Y - \widehat{f}(\boldsymbol{X})\right]^2\right\}, \tag{10.15}$$

where $\widehat{f}(\boldsymbol{X})$ is again the prediction for Y at a point $\boldsymbol{X}$, with $\boldsymbol{X}, Y$ drawn from their joint distribution.

Estimators $\widehat{f}$ with small $\text{ESPE}(\widehat{f})$ are sought, but balancing the bias in estimation with the variance will be a constant challenge. To illustrate, suppose we wish to

Table 10.2 Summary of predictive accuracy measures

Name	Short-hand	Definition
Generalization error	$\text{GE}(\widehat{f})$	$\text{E}_{\boldsymbol{X},Y}\left\{L[Y,\widehat{f}(\boldsymbol{X})]\right\}$
Expected squared prediction error	$\text{ESPE}(\widehat{f})$	$\text{E}_{\boldsymbol{X},Y}\left\{[Y-\widehat{f}(\boldsymbol{X})]^2\right\}$
Mean squared error (or risk)	$\text{MSE}\left[\widehat{f}(\boldsymbol{x}_0)\right]$	$\text{E}_{\boldsymbol{Y}_n}\left\{[\widehat{f}(\boldsymbol{x}_0)-f(\boldsymbol{x}_0)]^2\right\}$
Predictive risk	$\text{PR}\left[\widehat{f}(\boldsymbol{x}_0)\right]$	$\text{E}_{\boldsymbol{Y}_n,Y_0}\left\{[Y_0-\widehat{f}(\boldsymbol{x}_0)]^2\right\}$ $=\sigma^2+\text{MSE}\left[\widehat{f}(\boldsymbol{x}_0)\right]$
Integrated mean squared error	$\text{IMSE}\left(\widehat{f}\right)$	$\int \text{E}_{\boldsymbol{Y}_n}\left\{[\widehat{f}(\boldsymbol{x})-f(\boldsymbol{x})]^2\right\}p(\boldsymbol{x})\,d\boldsymbol{x}$ $=\int \text{MSE}\left(\widehat{f}(\boldsymbol{x})\right)p(\boldsymbol{x})\,d\boldsymbol{x}$
Average mean squared error	$\text{AMSE}\left(\widehat{f}\right)$	$n^{-1}\sum_{i=1}^n \text{e}_{\boldsymbol{Y}_n}\left\{[\widehat{f}(\boldsymbol{x}_i)-f(\boldsymbol{x}_i)]^2\right\}$ $=\sum_{i=1}^n \text{MSE}\left[\widehat{f}(\boldsymbol{x}_i)\right]$
Average predictive risk	$\text{APR}\left(\widehat{f}\right)$	$n^{-1}\sum_{i=1}^n \text{e}_{\boldsymbol{Y}_n,\boldsymbol{Y}_n^\star}\left\{[Y_i^\star-\widehat{f}(\boldsymbol{x}_i)]^2\right\}$ $=\sigma^2+\text{AMSE}\left(\widehat{f}\right)$
Residual sum of squares	$\text{RSS}\left(\widehat{f}\right)$	$n^{-1}\sum_{i=1}^n[y_i-\widehat{f}(\boldsymbol{x}_i)]^2$
Leave-one-out (ordinary) CV score	$\text{OCV}\left(\widehat{f}\right)$	$n^{-1}\sum_{i=1}^n[y_i-\widehat{f}_{-i}(\boldsymbol{x}_i)]^2$
Generalized CV score	$\text{GCV}\left[\widehat{f}\right]$	$[n-\text{tr}(\boldsymbol{S})]^{-1}\sum_{i=1}^n[y_i-\widehat{f}(\boldsymbol{x}_i)]^2$

All rows of the table but the first are based on integrated or summed *squared* quantities and, hence, are appropriate for a model of the form $y = f(\boldsymbol{x}) + \epsilon$ with the error terms ϵ having zero mean, constant variance σ^2, and with error terms at different $\boldsymbol{x}$ being uncorrelated. CV is short for cross-validation, with OCV and GCV being described in Sects. 10.6.2 and 10.6.3, respectively. Notation: The predictive model evaluated at covariates $\boldsymbol{x}$ is $f(\boldsymbol{x})$, with prediction $\widehat{f}(\boldsymbol{x})$ based on the observed data $\boldsymbol{Y}_n = [Y_1,\dots,Y_n]$; Y_0 is a new response with associated covariates $\boldsymbol{x}_0$; the observed data are $[y_i,\boldsymbol{x}_i]$, $i = 1,\dots,n$; $\boldsymbol{Y}_n^\star = [Y_1^\star,\dots,Y_n^\star]$ are a set of new observations with covariates $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ that we would like to predict; $p(\boldsymbol{x})$ is the distribution of the covariates; $\widehat{f}_{-i}(\boldsymbol{x}_i)$ is the prediction at the point $\boldsymbol{x}_i$ based on the observed data with the i-th case, $[y_i,\boldsymbol{x}_i]$, removed; $\boldsymbol{S}$ is the "smoother" hat matrix and is described in Sect. 10.6.1. The entries in the last three lines are all estimates of $\text{ESPE}(\widehat{f})$

predict a response Y_0 with associated covariates $\boldsymbol{x}_0$. We calculate the expected squared distance between the response Y_0 and the fitted function $\widehat{f}(x_0)$. The expectation is with respect to both Y_0 and repeat (training) data $\boldsymbol{Y}_n = [Y_1,\dots,Y_n]$ with Y_0 and $\boldsymbol{Y}_n$ being independent. The resultant measure is known as the *predictive risk* and may be decomposed as

$$\begin{aligned}
\text{E}_{\boldsymbol{Y}_n,Y_0}\left\{\left[Y_0-\widehat{f}(\boldsymbol{x}_0)\right]^2\right\} &= \text{E}_{\boldsymbol{Y}_n,Y_0}\left\{\left[Y_0-f(\boldsymbol{x}_0)+f(\boldsymbol{x}_0)-\widehat{f}(\boldsymbol{x}_0)\right]^2\right\}\\
&= \text{E}_{Y_0}\left\{[Y_0-f(\boldsymbol{x}_0)]^2\right\}+\text{E}_{\boldsymbol{Y}_n}\left\{\left[\widehat{f}(\boldsymbol{x}_0)-f(\boldsymbol{x}_0)\right]^2\right\}\\
&\quad +2\times\text{E}_{Y_0}\left\{[Y_0-f(\boldsymbol{x}_0)]\right\}\text{E}_{\boldsymbol{Y}_n}\left\{\left[\widehat{f}(\boldsymbol{x}_0)-f(\boldsymbol{x}_0)\right]\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \sigma^2 + \mathrm{E}_{\boldsymbol{Y}_n}\left\{\left[\widehat{f}(\boldsymbol{x}_0) - f(\boldsymbol{x}_0)\right]^2\right\} \\
&= \sigma^2 + \mathrm{MSE}\left[\widehat{f}(\boldsymbol{x}_0)\right].
\end{aligned}
$$

Writing

$$
\mathrm{MSE}\left[\widehat{f}(\boldsymbol{x}_0)\right] = \mathrm{E}_{\boldsymbol{Y}_n}\left\{\left[f(\boldsymbol{x}_0) - \mathrm{E}_{\boldsymbol{Y}_n}\left(\widehat{f}(\boldsymbol{x}_0)\right) + \mathrm{E}_{\boldsymbol{Y}_n}\left(\widehat{f}(\boldsymbol{x}_0)\right) - \widehat{f}(\boldsymbol{x}_0)\right]^2\right\}
$$

we have

$$
\begin{aligned}
\mathrm{E}_{\boldsymbol{Y}_n, Y_0}\left\{\left[Y_0 - \widehat{f}(\boldsymbol{x}_0)\right]^2\right\} &= \sigma^2 + \mathrm{E}_{\boldsymbol{Y}_n}\left\{\left[\mathrm{E}_{\boldsymbol{Y}_n}\left(\widehat{f}(\boldsymbol{x}_0)\right) - f(\boldsymbol{x}_0)\right]^2\right\} \\
&\quad + \mathrm{E}_{\boldsymbol{Y}_n}\left\{\left[\widehat{f}(\boldsymbol{x}_0) - \mathrm{E}_{\boldsymbol{Y}_n}\left(\widehat{f}(\boldsymbol{x}_0)\right)\right]^2\right\} \\
&= \sigma^2 + \mathrm{bias}\left[\widehat{f}(\boldsymbol{x}_0)\right]^2 + \mathrm{var}_{\boldsymbol{Y}_n}\left[\widehat{f}(\boldsymbol{x}_0)\right].
\end{aligned}
$$

In terms of the prediction error we can achieve given a particular model, nothing can be done about σ^2, which is referred to as the *irreducible error*. Therefore, we concentrate on the MSE of the estimator $\widehat{f}(\boldsymbol{x}_0)$:

$$
\mathrm{MSE}\left[\widehat{f}(\boldsymbol{x}_0)\right] = \mathrm{E}_{\boldsymbol{Y}_n}\left\{\left[\widehat{f}(\boldsymbol{x}_0) - f(\boldsymbol{x}_0)\right]^2\right\} = \mathrm{bias}\left[\widehat{f}(\boldsymbol{x}_0)\right]^2 + \mathrm{var}\left[\widehat{f}(\boldsymbol{x}_0)\right]
$$

where we emphasize that the MSE is calculated at the point $\boldsymbol{x}_0$, with the expectation over training samples. As we discuss subsequently, the estimators $\widehat{f}$ we consider are indexed by a smoothing parameter, and selection of this parameter influences the characteristics of $\widehat{f}$. Little smoothing produces a wiggly $\widehat{f}$, with low bias and high variance. More extensive smoothing produces $\widehat{f}$ with greater bias but reduced variance.

To summarize the MSE over the range of $\boldsymbol{x}$, we may consider the *integrated mean squared error* (IMSE). For univariate x, over an interval $[a, b]$, and with density $p(x)$:

$$
\begin{aligned}
\mathrm{IMSE}\left(\widehat{f}\right) &= \int_a^b \mathrm{E}_{\boldsymbol{Y}_n}\left\{\left[\widehat{f}(x) - f(x)\right]^2\right\} p(x)\, dx \\
&= \int_a^b \mathrm{bias}\left[\widehat{f}(x)\right]^2 p(x)\, dx + \int_a^b \mathrm{var}\left[\widehat{f}(x)\right] p(x)\, dx.
\end{aligned}
\tag{10.16}
$$

This summary will be encountered in Sect. 11.3.2.

An alternative to the IMSE, that may be more convenient to use, is the *average mean squared error* (AMSE), which only considers the errors at the observations:

$$\begin{aligned}\text{AMSE}\left(\widehat{f}\right) &= \frac{1}{n}\sum_{i=1}^{n} \text{E}_{\boldsymbol{Y}_n}\left\{\left[\widehat{f}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)\right]^2\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} \text{bias}\left[\widehat{f}(\boldsymbol{x}_i)\right]^2 + \frac{1}{n}\sum_{i=1}^{n} \text{var}\left[\widehat{f}(\boldsymbol{x}_i)\right]. \qquad (10.17)\end{aligned}$$

For the additive errors model (10.13), the *average predictive risk* (APR) is

$$\begin{aligned}\text{APR}\left(\widehat{f}\right) &= \frac{1}{n}\sum_{i=1}^{n} \text{E}_{\boldsymbol{Y}_n,\boldsymbol{Y}_n^\star}\left\{\left[Y_i^\star - \widehat{f}(\boldsymbol{x}_i)\right]^2\right\} \\ &= \sigma^2 + \text{AMSE}\left(\widehat{f}\right).\end{aligned}$$

where $\boldsymbol{Y}_n^\star = [Y_1^\star, \ldots, Y_n^\star]$ are the new set of observations which we would like to predict at $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, and are independent of $\boldsymbol{Y}_n$. In Sect. 10.6.1, a procedure for estimating the APR will be described in the context of smoothing parameter choice.

We denote the test data by $[y_i^\star, \boldsymbol{x}_i^\star]$, $i = 1, \ldots, m$. For continuous data and quadratic loss, we may evaluate an estimate of the expected squared prediction error (10.15):

$$\frac{1}{m}\sum_{i=1}^{m}\left[y_i^\star - \widehat{f}(\boldsymbol{x}_i^\star)\right]^2, \qquad (10.18)$$

where $\widehat{f}(\boldsymbol{x}_i^\star)$ is the estimator based on the training data.

10.4.2 Discrete Responses with K Categories

With the loss function (10.7), and with equal losses, the generalization error is

$$\Pr_{X,\,Y}\left[\widehat{g}(\boldsymbol{X}) \neq Y\right], \qquad (10.19)$$

which is also known as the *misclassification probability*. Given test data $[y_i^\star, \boldsymbol{x}_i^\star]$, $i = 1, \ldots, m$, the empirical estimate is

$$\frac{1}{m}\sum_{i=1}^{m} I\left[\widehat{g}(\boldsymbol{x}_i^\star) \neq y_i^\star\right],$$

which is simply the proportion of misclassified observations.

We now consider the binary case and introduce terminology that is common in a medical context, before describing additional measures that are useful summaries of a procedure in this case. Suppose we wish to predict disease status given covariates (symptoms) $\boldsymbol{x}$. Define

$$Y = \begin{cases} 0 & \text{if true state is no disease} \\ 1 & \text{if true state is disease.} \end{cases}$$

A classification rule $g(\boldsymbol{x})$ is

$$g(\boldsymbol{x}) = \begin{cases} 0 & \text{if prediction is no disease} \\ 1 & \text{if prediction is disease.} \end{cases}$$

The *sensitivity* of a rule is the probability of predicting disease for a diseased individual:

$$\text{Sensitivity} = \Pr\left[g(\boldsymbol{x}) = 1 \mid Y = 1\right].$$

The *specificity* is the probability of predicting disease-free for an individual without disease:

$$\text{Specificity} = \Pr\left[g(\boldsymbol{x}) = 0 \mid Y = 0\right].$$

With respect to Table 10.1, recall that $L(0,1)$ is the loss for predicting $g(\boldsymbol{x}) = 1$ when in reality $Y = 0$ (so we predict disease for a healthy individual) and $L(1,0)$ is the loss associated with predicting healthy for a diseased individual. Consequently, if we increase the former loss $L(0,1)$ while holding $L(1,0)$ constant, we will be more conservative in declaring a patient as diseased, which will increase the specificity and decrease the sensitivity.[4] An alternative, closely related, pair of summaries are the *false-positive fraction* (FPF) and *true-positive fraction* (TPF) defined, respectively, as

$$\text{FPF} = \Pr\left[g(\boldsymbol{X}) = 1 \mid Y = 0\right]$$

and

$$\text{TPF} = \Pr\left[g(\boldsymbol{X}) = 1 \mid Y = 1\right].$$

The sensitivity is the TPF, and the specificity is $(1-\text{FPF})$. Two additional measures are the *positive predictive value* (PPV) and the *negative predictive value* (NPV), defined as

$$\text{PPV} = \Pr\left[Y = 1 \mid g(\boldsymbol{X}) = 1\right] = \frac{\Pr\left[g(\boldsymbol{X}) = 1 \mid Y = 1\right]\Pr(Y = 1)}{\Pr\left[g(\boldsymbol{X}) = 1\right]}$$

$$\text{NPV} = \Pr\left[Y = 0 \mid g(\boldsymbol{X}) = 0\right] = \frac{\Pr\left[g(\boldsymbol{X}) = 0 \mid Y = 0\right]\Pr(Y = 0)}{\Pr\left[g(\boldsymbol{X}) = 0\right]},$$

which give the probabilities of correct assignments, given classification.

[4] We note that the decision problem considered here has many elements in common with that in which we choose between two hypotheses, as discussed in Sect. 4.3.1. The sensitivity is analogous to the power of a test, while $1-$specificity is analogous to the type I error.

Now define a classification rule that, based on a model $g(\boldsymbol{x})$ (whose parameters will be estimated from the data), assigns $g(\boldsymbol{x}) = 1$ if the odds of disease

$$\frac{\Pr(Y = 1 \mid \boldsymbol{x})}{\Pr(Y = 0 \mid \boldsymbol{x})} > \frac{L(0, 1)}{L(1, 0)} = R,$$

as discussed in more detail in relation to (10.9). Plotting $\text{TPF}(R)$ versus $\text{FPF}(R)$ produces a *receiver-operating characteristic* (ROC) curve. The ROC curve gives the complete behavior of FPF and TPF over the range of R. Pepe (2003) provides an in-depth discussion of the above summary measures.

10.4.3 General Responses

For general data types we may evaluate the deviance-like loss function (10.12) over the test data $[y_i^\star, \boldsymbol{x}_i^\star]$, $i = 1, \ldots, m$:

$$-\frac{2}{m} \sum_{i=1}^{m} \log p_{\widehat{f}}\left(y_i^\star \mid \boldsymbol{x}_i^\star\right),$$

to measure the error of a procedure.

10.5 A First Look at Shrinkage Methods

We describe two penalization methods that are used in the context of multiple linear regression, ridge regression and the lasso.

10.5.1 Ridge Regression

We first assume that $\boldsymbol{y}$ has been centered and that each covariate has been standardized, that is,

$$\sum_{i=1}^{n} y_i = 0, \quad \frac{1}{n} \sum_{i=1}^{n} x_{ij} = 0, \quad \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 = 1.$$

Consider the linear model

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with $\boldsymbol{x}$ the $n \times k$ design matrix, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_k]^{\mathrm{T}}$ the $k \times 1$ vector of parameters, and $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$. Note that there is no intercept in the model due to the centering of $y_1, \ldots, y_n$.

We saw in Chap. 5 that linear models are an analytically and computationally appealing class but, with many predictors, fitting the full model without penalization may result in large predictive intervals, unless the sample size is very large relative to k. Ridge regression is an approach to modeling that addresses this deficiency by placing a particular form of constraint on the parameters. Specifically, $\widehat{\boldsymbol{\beta}}^{\text{RIDGE}}$ is chosen to minimize the *penalized sum of squares*:

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{k} \beta_j^2, \tag{10.20}$$

for some $\lambda > 0$. Using a Lagrange multiplier argument (Exercise 10.6), minimization of (10.20) is equivalent to minimization of

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} x_{ij} \beta_j \right)^2$$

subject to, for some $s \geq 0$,

$$\sum_{j=1}^{k} \beta_j^2 \leq s, \tag{10.21}$$

so that the size of the sum of the squared coefficients is constrained (which is known as an L_2 penalty). The intuition behind ridge regression is that, with many parameters to estimate, the estimator can be highly variable, but by constraining the sum of the squared coefficients, this shortcoming can be alleviated.

Figure 10.3 shows the effect of ridge regression with two parameters, β_1 and β_2. The elliptical contours in the top right of the figure correspond to the sum of squares. In ridge regression this sum of squares is minimized subject to the constraint (10.21), and for $k = 2$, this constraint corresponds to a circle, centered at zero. The estimate is given by the point at which the ellipse and the circle touch.

Writing the penalized sum of squares (10.20) as

$$(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\beta} \tag{10.22}$$

it is easy to see that the minimizing solution is

$$\widehat{\boldsymbol{\beta}}^{\text{RIDGE}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \lambda \mathbf{I}_k)^{-1} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{Y}. \tag{10.23}$$

Since the estimator (10.23) is linear, it is straightforward to calculate the variance–covariance matrix, for a given λ, as

$$\mathrm{var}\left(\widehat{\boldsymbol{\beta}}^{\text{RIDGE}} \right) = \sigma^2 (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \lambda \mathbf{I}_k)^{-1} \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \lambda \mathbf{I}_k)^{-1}. \tag{10.24}$$

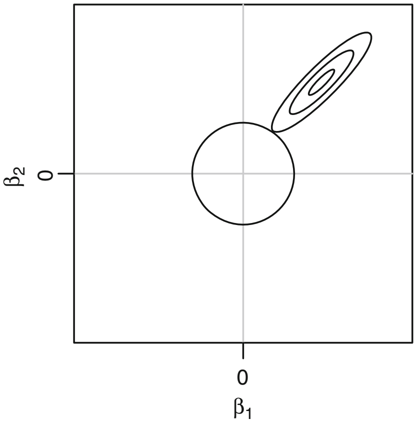


Fig. 10.3 Pictorial representation of ridge regression, for two covariates. The *elliptical contours* represent the sum of squares, and the *circle* represents the constraint corresponding to the L_2 penalty

Beginning with a normal likelihood $\boldsymbol{y} \mid \boldsymbol{\beta} \sim \mathrm{N}_n(\boldsymbol{x\beta}, \sigma^2\mathbf{I}_n)$ and adding the penalty term $\lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}$ to the log-likelihood also leads to minimization of (10.22). The resultant estimator (10.23) is therefore sometimes referred to as a *maximum penalized likelihood estimator* (MPLE).

It is well known that the least squares estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$ is an unbiased estimator, with variance $(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\sigma^2$ (under correct second moment specification). If we write $\boldsymbol{R} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}$, then the ridge regression estimator may be written as (Exercise 10.7)

$$\widehat{\boldsymbol{\beta}}^{\mathrm{RIDGE}} = (\mathbf{I}_k + \lambda\boldsymbol{R})^{-1}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}, \tag{10.25}$$

showing that it is clearly biased. Turning now to a consideration of the variance, let $\boldsymbol{x} = \boldsymbol{UDV}^{\mathrm{T}}$ be the singular value decomposition (SVD) of $\boldsymbol{x}$. In the SVD $\boldsymbol{U}$ is $n \times n$, $\boldsymbol{V}$ is $k \times k$ and $\boldsymbol{D}$ is an $n \times k$ diagonal matrix with diagonal elements $d_1, \ldots, d_k$. Then, the variance of the ridge estimator (10.24) may be written as

$$\mathrm{var}\left(\widehat{\boldsymbol{\beta}}^{\mathrm{RIDGE}}\right) = \sigma^2(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \lambda\mathbf{I}_k)^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \lambda\mathbf{I}_k)^{-1} = \sigma^2\boldsymbol{VAV}^{\mathrm{T}}, \tag{10.26}$$

where $\boldsymbol{A}$ is a diagonal matrix whose elements are $d_i^2/(d_i^2 + \lambda)^2$. The variance of the least squares estimator is

$$\mathrm{var}\left(\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}\right) = \sigma^2\boldsymbol{VWV}^{\mathrm{T}}, \tag{10.27}$$

where $\boldsymbol{W}$ is a diagonal matrix whose elements are $1/d_i^2$. Hence, the reduction in variance of the ridge regression estimator is apparent. The derivations of (10.26) and (10.27) are left as Exercise 10.7.

With respect to the frequentist methods described in Chap. 2, penalized least squares correspond to a method that produces an estimating function with finite sample bias but with potentially lower mean squared error as a consequence of the penalization term, which reduces the variance.

For *orthogonal* covariates $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = n \times \mathbf{I}_k$, the ridge regression estimator is

$$\widehat{\boldsymbol{\beta}}^{\text{RIDGE}} = \frac{n}{n+\lambda}\widehat{\boldsymbol{\beta}}^{\text{LS}}.$$

Hence, in this case, the ridge estimator always produces shrinkage towards 0. Figure 10.4(a) illustrates the shrinkage (towards zero) performed by ridge regression for a single parameter in the case of orthogonal covariates. For non-orthogonal covariates, the collection of estimators undergoes shrinkage, though individual components of $\widehat{\boldsymbol{\beta}}^{\text{RIDGE}}$ may increase in absolute value.

The fitted value at a particular value $\widetilde{\boldsymbol{x}}$ is

$$\widehat{f}\,(\,\widetilde{\boldsymbol{x}}\,) = \widetilde{\boldsymbol{x}}\,\widehat{\boldsymbol{\beta}}^{\text{RIDGE}} \tag{10.28}$$

$$= \widetilde{\boldsymbol{x}}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \lambda\mathbf{I}_k)^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Y} \tag{10.29}$$

with

$$\text{var}\left[\widehat{f}\,(\,\widetilde{\boldsymbol{x}}\,)\right] = \sigma^2\widetilde{\boldsymbol{x}}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \lambda\mathbf{I}_k)^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \lambda\mathbf{I}_k)^{-1}\widetilde{\boldsymbol{x}}^{\mathrm{T}}. \tag{10.30}$$

An important concept in shrinkage is the “effective” degrees of freedom associated with a set of parameters. In a ridge regression setting, if we choose $\lambda = 0$, we have k parameters, while for $\lambda > 0$ the parameters are constrained and the degrees of freedom will effectively be lower, tending to 0 as $\lambda \to \infty$. Many smoothers are linear in the sense that $\widehat{y} = \boldsymbol{S}^{(\lambda)}y$, with ridge regression being one example, as can be seen from (10.29). For linear smoothers, the *effective (or equivalent) degrees of freedom* may be defined as

$$p^{(\lambda)} = \text{df}(\lambda) = \text{tr}\left(\boldsymbol{S}^{(\lambda)}\right), \tag{10.31}$$

where the notation $p^{(\lambda)}$ emphasizes the dependence on the smoothing parameter. For the ridge estimator, the effective degrees of freedom associated with estimation of $\beta_1, \ldots, \beta_k$ is defined as

$$\text{df}(\lambda) = \text{tr}\left[\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \lambda\mathbf{I}_k)^{-1}\boldsymbol{x}^{\mathrm{T}}\right]. \tag{10.32}$$

Notice that $\lambda = 0$, which corresponds to no shrinkage, gives $\text{df}(\lambda) = k$ (so long as $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$ is non-singular), as we would expect.

There is a one-to-one mapping between λ and the degrees of freedom, so in practice, one may simply pick the effective degrees of freedom that one would like associated with the fit and solve for λ. As an alternative to a user-chosen λ, a number of automated methods for choosing λ are described in Sect. 10.6.

Insight into the ridge estimator can be gleaned from the following Bayesian formulation. Consider the model with likelihood

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2 \sim \text{N}_n(\boldsymbol{x}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \tag{10.33}$$

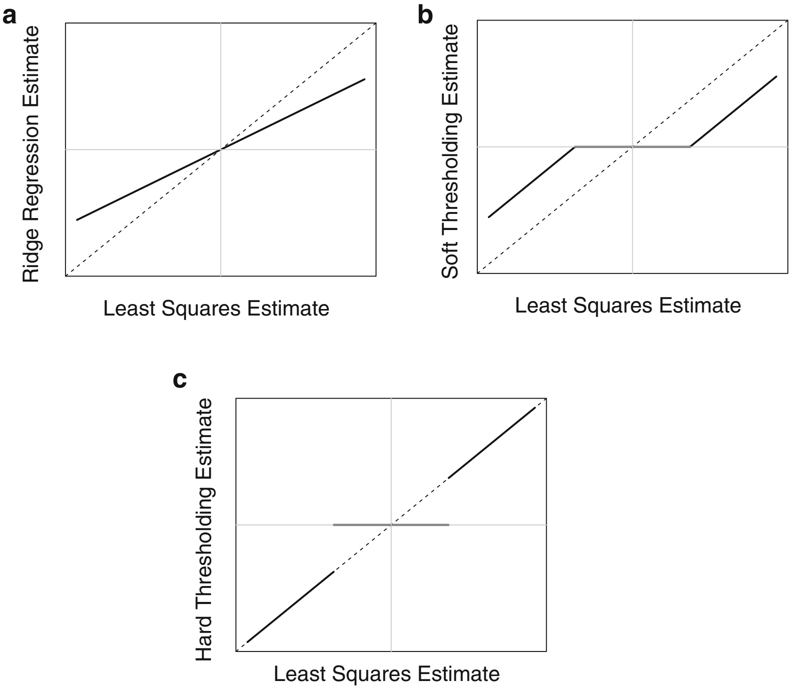


Fig. 10.4 The comparison for single estimate of different forms of shrinkage, with alternative estimates plotted against the least squares estimate $\widehat{\beta}^{\text{LS}}$ and in the case of orthogonal covariates: (**a**) ridge regression, (**b**) *soft thresholding* as carried out by the lasso, and (**c**) *hard thresholding* as carried out by conventional variable selection. On all plots, the line of equality, representing the unrestricted estimate, is drawn as *dashed*

with σ^2 known, and prior

$$\boldsymbol{\beta} \mid \sigma^2 \sim \mathrm{N}_k \left(\mathbf{0}, \frac{\sigma^2}{\lambda} \mathbf{I}_k \right).$$

The latter form shows that a large value of λ corresponds to a prior that is more tightly concentrated around zero and so leads to greater *shrinkage* of the collection of coefficients towards zero. A common λ for each β_j makes it clear that we need to standardize each of the covariates in order for them to be comparable.

Using derivations similar to those of Sect. 5.7, the posterior is

$$\boldsymbol{\beta} \mid \boldsymbol{y} \sim \mathrm{N}_k \left[\widehat{\boldsymbol{\beta}}^{\text{RIDGE}}, \sigma^2 (\boldsymbol{x}^{\mathrm{T}} \boldsymbol{x} + \lambda \mathbf{I}_k)^{-1} \right],$$

where $\widehat{\boldsymbol{\beta}}^{\text{RIDGE}}$ corresponds to (10.23), confirming that the posterior mean and mode coincide with the ridge regression estimator, (10.23). Interestingly, the posterior variance $\text{var}(\boldsymbol{\beta} \mid \boldsymbol{y})$ differs from var $\left(\widehat{\boldsymbol{\beta}}^{\text{RIDGE}}\right)$, as given in (10.24).

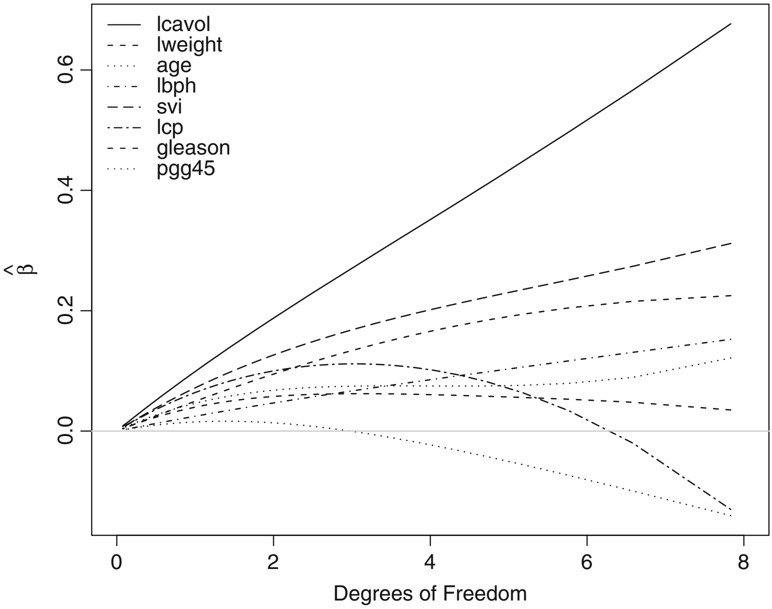


Fig. 10.5 Ridge estimates for the prostate data, as a function of the effective degrees of freedom

Example: Prostate Cancer

As described in Sect. 1.3.1, the response in this dataset is log (PSA), and there are eight covariates. In this chapter, we take the aim of the analysis as prediction of log PSA. In Chap. 5, we analyzed these data using a Bayesian approach with normal priors for each of the eight standardized coefficients, as summarized in (5.66). In that case, the standard deviation of the normal prior was chosen on substantive grounds. Here, we illustrate the behavior of the estimates as a function of the smoothing parameter.

Figure 10.5 shows the eight ridge estimates as a function of the effective degrees of freedom (which ranges between 0 and 8, because there is no intercept in the model). For small values of λ, the effective degrees of freedom is close to 8, and estimates show little shrinkage. In contrast, large values of λ give effective degrees of freedom close to 0 and strong shrinkage. Notice that the curves do not display monotonic shrinkage due to the non-orthogonality of the covariates.

10.5.2 The Lasso

The *least absolute shrinkage and selection operator*, or *lasso*, as described in Tibshirani (1996),[5] is a technique that has received a great deal of interest. As with ridge regression, we assume that the covariates are standardized to have mean zero and standard deviation 1. The lasso estimate minimizes the penalized sum of squares

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{k} |\beta_j|, \tag{10.34}$$

with respect to $\boldsymbol{\beta}$. The L_2 penalty of ridge regression is therefore being replaced by an L_1 penalty. As with ridge regression, the minimization of (10.34) can be shown to be equivalent to minimization of

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2 \tag{10.35}$$

subject to

$$\sum_{j=1}^{k} |\beta_j| \leq s, \tag{10.36}$$

for some $s \geq 0$.

Let $\widehat{\boldsymbol{\beta}}^{\text{LS}}$ and $\widehat{\boldsymbol{\beta}}^{\text{LASSO}}$ denote the least squares and lasso estimates, respectively, and define $s_0 = \sum_{j=1}^{k} |\widehat{\beta}_j^{\text{LS}}|$ as the L_1 norm of the least squares estimate. Values of $s < s_0$ cause shrinkage of $\sum_{j=1}^{k} |\widehat{\beta}_j^{\text{LASSO}}|$ towards zero. If, for example, $s = s_0/2$, then the average absolute shrinkage of the least squares coefficients is 50%, though individual coefficients may increase rather than decrease in absolute value.

A key characteristic of the lasso is that individual parameter estimates may be set to zero, a phenomenon that does not occur with ridge regression. Figure 10.6 gives the intuition behind this behavior in the case of two coefficients β_1 and β_2. The lasso performs L_1 shrinkage so that there are "corners" in the constraint; the diamond represents constraint (10.36) for $k = 2$. If the ellipse (10.35) "hits" one of these corners, then the coefficient corresponding to the axis that is touched is shrunk to zero. In the example in Fig. 10.6, neither of the coefficients would be set to zero, because the ellipse does not touch a corner. As k increases, the multidimensional diamond has an increasing number of corners, and so there is an increasing chance of coefficients being set to zero. Consequently, the lasso effectively produces a form of *continuous* subset (or feature) selection. The lasso is sometimes referred to as offering a *sparse solution* due to this property of setting coefficients to zero.

[5] The method was also introduced into the signal-processing literature, under the name *basis pursuit*, by Chen et al. (1998).

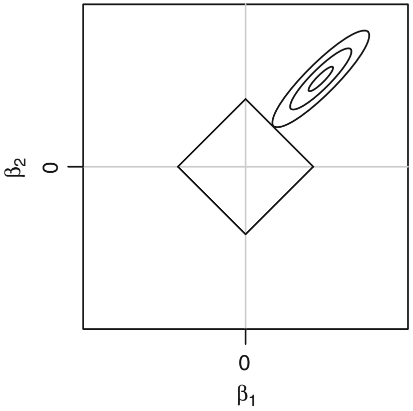


Fig. 10.6 Pictorial representation of the lasso for two covariates. The *elliptical contours* represent the sum of squares, and the *diamond* indicates the constraint corresponding to the L_1 penalty

In the case of orthonormal covariates, for which $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = \mathbf{I}_k$, the lasso performs so-called *soft thresholding*. Specifically, for component j of the lasso estimator:

$$\widehat{\beta}_j^{\text{LASSO}} = \text{sign}\left(\widehat{\beta}_j^{\text{LS}}\right)\left(|\widehat{\beta}_j^{\text{LS}}| - \frac{\lambda}{2}\right)_+,$$

where "sign" denotes the sign of its argument (±1), and z_+ represents the positive part of z. As the smoothing parameter is varied, the sample path of the estimates moves continuously to zero, as displayed in Fig. 10.4(b). In contrast, conventional hypothesis testing performs *hard thresholding*, as illustrated in Fig. 10.4(c), since the coefficient is set equal to zero when the absolute value of the estimate drops below some critical value, giving discontinuities in the graph.

The lasso solution is nonlinear in $\boldsymbol{y}$. Efficient algorithms exist for computation based on coordinate descent; however, see Meier et al. (2008) and Wu and Lange (2008). Tibshirani (2011) gives a brief history of the computation of the lasso solution. Due to the nonlinearity of the solution and the subset selection nature of estimation, inference is not straightforward and remains an open problem. Standard errors for elements of $\widehat{\boldsymbol{\beta}}^{\text{LASSO}}$ are not immediately available, though they may be calculated via the bootstrap. Since the lasso estimator is not linear, the effective degrees of freedom cannot be defined as in (10.31); an alternative definition exists as

$$\text{df} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\text{cov}(\widehat{y}_i, y_i),$$

see Hastie et al. (2009) equation (3.60).

More generally, penalties of the form

$$\lambda\sum_{j=1}^{k}|\beta_j|^q$$

may be considered, for $q \geq 0$. Ridge regression and the lasso correspond to $q = 2$ and $q = 1$, respectively. For $q < 1$, the constraint is non-convex, which makes optimization more difficult. Convex penalties occur for $q \geq 1$ and feature selection for $q \leq 1$, so that the lasso (with $q = 1$) achieves both.

Many variants of the lasso have appeared since its introduction (Tibshirani 2011). In some contexts, we may wish to treat a set of regressors as a group, for example, when we have a categorical covariate with more than two levels. The *grouped lasso* (Yuan and Lin 2007) addresses this problem by considering the simultaneous shrinkage of (pre-defined) groups of coefficients.

In the case in which $k > n$, the lasso cannot select more than n variables. Furthermore, the lasso will typically assign only one nonzero coefficient to a set of highly correlated covariates (Zou and Hastie 2005), which is an obvious disadvantage and was a motivation for the group lasso (Yuan and Lin 2007). Empirical observation indicates that the lasso produces inferior performance to ridge regression when there are a large number of small effects (Tibshirani 1996). These deficiencies motivated the *elastic net* (Zou and Hastie 2005) which attempts to combine the desirable properties of ridge regression and the lasso via a penalty of the form

$$\lambda_1 \sum_{j=1}^{k} |\beta_j| + \lambda_2 \sum_{j=1}^{k} \beta_j^2 .$$

The lasso estimate is equivalent to the mode of the posterior distribution under a normal likelihood, (10.33), and independent Laplace (double exponential) priors on elements of $\boldsymbol{\beta}$:

$$\pi(\beta_j) = \frac{\lambda}{2} \exp\left(-\lambda |\beta_j|\right)$$

for $j = 1, \ldots, k$ (the variance of this distribution is $2/\lambda^2$, Appendix D). Under this prior, the posterior is not available in closed form, but the posterior mean will not equal the posterior mode. Hence, if used as a summary, the posterior means will not produce the same lasso shrinkage of coefficients to zero. Thus, regardless of the value of λ, all k covariates are retained in a Bayesian analysis, even though the posterior mode may lie at zero. Markov chain Monte Carlo allows inference under the normal/Laplace model but without the subset selection aspect, which lessens the appeal of this Bayesian version of the lasso.

Example: Prostate Cancer

We illustrate the use of the lasso for the prostate cancer data. Figure 10.7 shows the lasso estimates as a function of the shrinkage factor:

$$\frac{\sum_{j=1}^{k} |\widehat{\beta}_j^{\text{LASSO}}|}{\sum_{j=1}^{k} |\widehat{\beta}_j^{\text{LS}}|} .$$

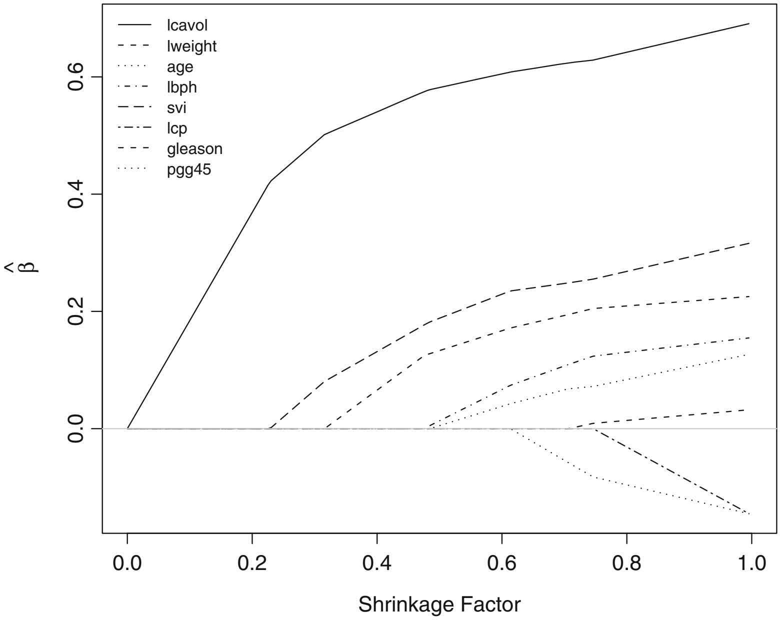


Fig. 10.7 Lasso estimates for the prostate data, as a function of the shrinkage factor, $\sum_{j=1}^{k} |\hat{\beta}_j^{\text{LASSO}}| / \sum_{j=1}^{k} |\hat{\beta}_j^{\text{LS}}|$

When the shrinkage factor is 1, the lasso estimates are the same as the least squares estimates. Beginning with the coefficient associated with log capsular penetration and ending with that associated with log cancer volume each of the coefficients is absorbed at zero, as the coefficient trajectories are traced out. For example, at a shrinkage factor of 0.4, only 3 coefficients are nonzero, those associated with log cancer volume, log weight and Gleason. In this example, the curves decrease monotonically to zero, but this phenomenon will not occur in all examples. The piecewise linear nature of the solution is apparent.

10.6 Smoothing Parameter Selection

For both ridge regression and the lasso, as well as a number of methods to be described in Chaps. 11 and 12, a key element of implementation is smoothing parameter selection.[6] We denote a generic smoothing parameter by λ and the estimated function at this λ, for a particular covariate value $\boldsymbol{x}$, by $\widehat{f}^{(\lambda)}(\boldsymbol{x})$.

[6] We use the name "smoothing" parameter because we concentrate on nonparametric regression smoothers in this and the next two chapters, but in the context of ridge regression and the lasso, the label "tuning" parameter is often used.

In this section, the overall strategy is to derive methods for minimizing, with respect to λ, estimates of the generalization error, or related measures. We initially assume a quadratic loss function before describing smoothing parameter selection in generalized linear model situations.

In Sect. 10.6.1, an analytic method of minimizing the AMSE (Table 10.2) is described and shown to be equivalent to Mallows C_P (Sect. 4.8.2). Two popular approaches for smoothing parameter selection, ordinary and generalized cross-validation, are described in Sects. 10.6.2 and 10.6.3, and in Sect. 10.6.4, we describe the AIC model selection statistic, which extends Mallows C_P to general data types. Finally, Sect. 10.6.5 briefly describes cross-validation for generalized linear models.

Bayesian approaches include choosing λ on substantive grounds (as carried out in Sect. 5.12) or treating λ as an unknown parameter. In the latter case, a prior is specified for λ, which is then estimated in the usual way. Section 11.2.8 adopts a mixed model formulation and describes a frequentist approach to smoothing parameter estimation, with restricted maximum likelihood (REML, see Sect. 8.5.3) being emphasized. Section 11.2.9 takes the same formulation but describes a Bayesian approaches to estimation.

Smoothing parameter choice is an inherently difficult problem because, in many situations, the data do not indicate a clear “optimal” λ. Therefore, there is no universally reliable method for smoothing parameter selection. Consequently, in practice, one should not blindly accept the solution provided by any method. Rather, one should treat the solution as a starting point for further exploration, including the use of alternative methods.

10.6.1 Mallows C_P

In this section we assume that the smoothing method produces a *linear smoother* of the form $\widehat{\boldsymbol{y}} = \boldsymbol{S}^{(\lambda)}\boldsymbol{y}$. Ridge regression provides an example with $\boldsymbol{S}^{(\lambda)} = \boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \lambda\mathbf{I}_k)^{-1}\boldsymbol{x}^{\mathrm{T}}$; the lasso does not fall within this class. Many methods that we describe in Chap. 11 produce smoothers of linear form.

Recall, from Sect. 5.11.2, that in linear regression $\widehat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$ where $\boldsymbol{S} = \boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}$ is the *hat* matrix, and $\mathrm{tr}(\boldsymbol{S})$ is both the number of regression parameters in the model and the degrees of freedom. Equation (10.31) defined the effective degrees of freedom for linear smoothers as $p^{(\lambda)} = \mathrm{df}(\lambda) = \mathrm{tr}\left(\boldsymbol{S}^{(\lambda)}\right)$. One approach to smoothing parameter choice is to simply pick λ to produce the desired effective degrees of freedom $p^{(\lambda)}$, if we have some a priori sense of the degrees of freedom that is desirable. This allows a direct comparison with parametric models. For example, one may pick $p^{(\lambda)} = 4$ to provide a fit with effective degrees of freedom equal to the number of parameters in a cubic polynomial regression model.

An appealing approach is to choose the smoothing parameter to minimize the average mean squared error, (10.17):

$$\begin{aligned}\text{AMSE}^{(\lambda)} = \text{AMSE}(\widehat{\boldsymbol{f}}^{(\lambda)}) &= \frac{1}{n}\sum_{i=1}^{n} \text{E}\left\{\left[f(\boldsymbol{x}_i) - \widehat{f}^{(\lambda)}(\boldsymbol{x}_i)\right]^2\right\} \\ &= \frac{1}{n}\text{E}\left[\left(\boldsymbol{f} - \widehat{\boldsymbol{f}}^{(\lambda)}\right)^{\mathrm{T}}\left(\boldsymbol{f} - \widehat{\boldsymbol{f}}^{(\lambda)}\right)\right], \qquad (10.37)\end{aligned}$$

where $\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]^{\mathrm{T}}$ and $\widehat{\boldsymbol{f}}^{(\lambda)} = \left[\widehat{f}^{(\lambda)}(\boldsymbol{x}_1), \ldots, \widehat{f}^{(\lambda)}(\boldsymbol{x}_n)\right]^{\mathrm{T}}$. The AMSE depends on the unknown $\boldsymbol{f}$ and so is not directly of use. A more applicable version is obtained by replacing $\boldsymbol{f}$ by $\boldsymbol{Y} - \boldsymbol{\epsilon}$ (with $\text{E}[\boldsymbol{\epsilon}] = 0$) and taking $\widehat{\boldsymbol{f}}^{(\lambda)} = \boldsymbol{S}^{(\lambda)}\boldsymbol{Y}$ to give

$$\begin{aligned}\text{AMSE}^{(\lambda)} &= \frac{1}{n}\text{E}\left[\left(\boldsymbol{Y} - \boldsymbol{\epsilon} - \boldsymbol{S}^{(\lambda)}\boldsymbol{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \boldsymbol{\epsilon} - \boldsymbol{S}^{(\lambda)}\boldsymbol{Y}\right)\right] \\ &= \frac{1}{n}\text{E}\left[\left(\boldsymbol{Y} - \boldsymbol{S}^{(\lambda)}\boldsymbol{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \boldsymbol{S}^{(\lambda)}\boldsymbol{Y}\right)\right] \\ &\quad + \frac{1}{n}\text{E}[\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}] - \frac{1}{n}\text{E}\left[2\boldsymbol{\epsilon}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{S}^{(\lambda)})\boldsymbol{Y}\right].\end{aligned}$$

Replacing $\boldsymbol{Y}$ by $\boldsymbol{f} + \boldsymbol{\epsilon}$ in the final term and rearranging gives

$$\begin{aligned}\text{AMSE}^{(\lambda)} &= \frac{1}{n}\text{E}\left[\left(\boldsymbol{Y} - \boldsymbol{S}^{(\lambda)}\boldsymbol{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \boldsymbol{S}^{(\lambda)}\boldsymbol{Y}\right)\right] \\ &\quad - \frac{1}{n}\text{E}\left[\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{f} - 2\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{S}^{(\lambda)}\boldsymbol{f} - 2\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{S}^{(\lambda)}\boldsymbol{\epsilon}\right].\end{aligned}$$

Since

$$\begin{aligned}\text{E}\left[\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{S}^{(\lambda)}\boldsymbol{\epsilon}\right] &= \text{E}\left[\text{tr}\left(\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{S}^{(\lambda)}\boldsymbol{\epsilon}\right)\right] = \text{E}\left[\text{tr}\left(\boldsymbol{S}^{(\lambda)}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\right)\right] = \text{tr}\left(\boldsymbol{S}^{(\lambda)}\boldsymbol{I}\sigma^2\right) = \sigma^2\text{tr}\left(\boldsymbol{S}^{(\lambda)}\right) \\ &= \sigma^2 p^{(\lambda)},\end{aligned}$$

and $\text{E}[2\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{f}] = \text{E}[2\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{S}^{(\lambda)}\boldsymbol{f}] = 0$, we obtain

$$\begin{aligned}\text{AMSE}^{(\lambda)} &= \frac{1}{n}\text{E}\left[\left(\boldsymbol{Y} - \boldsymbol{S}^{(\lambda)}\boldsymbol{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \boldsymbol{S}^{(\lambda)}\boldsymbol{Y}\right)\right] - \sigma^2 + \frac{2}{n}\text{E}\left[\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{S}^{(\lambda)}\boldsymbol{\epsilon}\right] \\ &= \frac{1}{n}\text{E}\left[\left(\boldsymbol{Y} - \boldsymbol{S}^{(\lambda)}\boldsymbol{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \boldsymbol{S}^{(\lambda)}\boldsymbol{Y}\right)\right] - \sigma^2 + \frac{2}{n}p^{(\lambda)}\sigma^2. \qquad (10.38)\end{aligned}$$

The natural estimator of (10.38) is

$$\begin{aligned}\widehat{\text{AMSE}}^{(\lambda)} &= \frac{1}{n}\left(\boldsymbol{Y} - \boldsymbol{S}^{(\lambda)}\boldsymbol{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \boldsymbol{S}^{(\lambda)}\boldsymbol{Y}\right) - \widehat{\sigma}^2_{\max} + \frac{2}{n}p^{(\lambda)}\widehat{\sigma}^2_{\max} \\ &= \frac{\widehat{\sigma}^2_{\max}}{n}\left[\frac{\text{RSS}^{(\lambda)}}{\widehat{\sigma}^2_{\max}} - \left(n - 2p^{(\lambda)}\right)\right]\end{aligned}$$

where $\widehat{\sigma}^2_{\max} > 0$ is an estimate from a maximal model (e.g., the full model in a regression setting). Minimizing the estimated AMSE$^{(\lambda)}$ as a function of λ is therefore equivalent to minimization of Mallows C_P statistic, (4.25):

$$\frac{\text{RSS}^{(\lambda)}}{\widehat{\sigma}^2_{\max}} - \left(n - 2p^{(\lambda)}\right). \tag{10.39}$$

A useful quantity to evaluate is the *average predictive risk* (APR, Table 10.2), which is the predictive risk at the observed $\boldsymbol{x}_i$, $i = 1, \dots, n$. Specifically,

$$\text{APR} = \sigma^2 + \text{AMSE}, \tag{10.40}$$

which can be estimated by

$$\begin{aligned}\widehat{\text{APR}}^{(\lambda)} &= \widehat{\sigma}^2_{\max} + \frac{1}{n}\text{RSS}^{(\lambda)} - \frac{1}{n}\left(n - 2p^{(\lambda)}\right)\widehat{\sigma}^2_{\max} \\ &= \frac{\text{RSS}^{(\lambda)}}{n} + \frac{2p^{(\lambda)}}{n}\widehat{\sigma}^2_{\max}. \end{aligned} \tag{10.41}$$

Estimating APR by the average residual sum of squares (i.e. the first term in (10.41)) is clearly subject to *overfitting* (and hence will be an underestimate), but this is corrected for by the second term.

10.6.2 *K*-Fold Cross-Validation

A widely used and simple method for estimating prediction error, and hence smoothing parameters, is cross-validation. If we try to estimate the APR, as given by (10.40), from the data directly, that is, using

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \widehat{f}^{(\lambda)}(\boldsymbol{x}_i)\right]^2 = \frac{\text{RSS}^{(\lambda)}}{n},$$

we will obtain an optimistic estimate because the data have been used twice: once to fit the model and once to estimate the predictive risk, as we saw in (10.41). The problem is that the idiosyncrasies of the particular realization of the data will influence coefficient estimates so that the model will, in turn, predict the data "too well". As noted in Sect. 10.4, ideally one would split the data to produce a validation dataset, with estimation of the generalization error being performed using the validation data. Unfortunately there are frequently insufficient data to carry out this step. However, cross-validation provides an approach in the same spirit to estimate the APR.

In K-fold validation, a fraction $(K-1)/K$ of the data are used to fit the model. The remaining fraction, $1/K$, are predicted, and these data are used to produce an estimate of the predictive risk. Let $\boldsymbol{y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_K]$ represent a particular

K-fold split of the $n \times 1$ data vector $\boldsymbol{y}$. Further, let $J(k)$ be the set of elements of $\{1, 2, \ldots, n\}$ that correspond to the indices of data points within split k, with $n_k = |J(k)|$ representing the cardinality of set k. Let $\boldsymbol{y}_{-k}$ be the data with the portion $\boldsymbol{y}_k$ removed and $\widehat{f}_{-k}^{(\lambda)}(\boldsymbol{x}_i)$ represent the i-th fitted value, computed from fitting a model using $\boldsymbol{y}_{-k}$. Cross-validation proceeds by cycling over $k = 1, \ldots, K$ through the following two steps:

1. Fit the model using $\boldsymbol{y}_{-k}$.
2. Use the fitted model to obtain predictions for the removed data, $\boldsymbol{y}_k$, and estimate the error as

$$\text{CV}_k^{(\lambda)} = \frac{1}{n_k} \sum_{i \in J(k)} \left[y_i - \widehat{f}_{-k}^{(\lambda)}(\boldsymbol{x}_i) \right]^2 . \tag{10.42}$$

The K prediction errors are averaged to give

$$\text{CV}^{(\lambda)} = \frac{1}{K} \sum_{k=1}^{K} \text{CV}_k^{(\lambda)} .$$

This procedure is repeated for each potential value of the smoothing parameter, λ. We emphasize that the data are split into K pieces *once*, and so the resultant datasets are the same across all λ.

Typical choices for K include 5, 10, and n, the latter being known as *leave-one-out* or *ordinary* cross-validation (OCV). Picking $K = n$ produces an estimate of the expected prediction error with the least bias, but this estimate can have high variance because the n training sets are so similar to one another. The computational burden of OCV can be heavy, though for a large class of smoothers this burden can be side-stepped, as we describe shortly. For smaller values of K, the variance of the expected prediction error estimator is smaller but there is greater bias. Breiman and Spector (1992) provide some discussion on choice of K and recommend $K = 5$ based on simulations in which the aim was subset selection. A number of authors (e.g., Hastie et al. 2009) routinely create an estimate of the standard error of the cross-validation score, (10.42). This estimate assumes independence of $\text{CV}_k^{(\lambda)}$, $k = 1, \ldots, K$, which is clearly not true since each pair of splits share a proportion $1 - 1/(K - 1)$ of the data.

We consider leave-one-out cross-validation in more detail. It would appear that we need to fit the model n times, but we show that, for a particular class of smoothers (to be described below),

$$\frac{1}{n} \sum_{i=1}^{n} \left[y_i - \widehat{f}_{-i}^{(\lambda)}(\boldsymbol{x}_i) \right]^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{\left[y_i - \widehat{f}^{(\lambda)}(\boldsymbol{x}_i) \right]^2}{\left(1 - S_{ii}^{(\lambda)} \right)^2} \tag{10.43}$$

where $S_{ii}^{(\lambda)}$ is the ith diagonal element of $\boldsymbol{S}^{(\lambda)}$, and $\widehat{f}_{-i}^{(\lambda)}(\boldsymbol{x}_i)$ is the ith fitted point, based on $\boldsymbol{y}_{-i}$.

We prove (10.43), for a particular class of smoothers, based on the derivation in Wood (2006, Sect. 4.5.2). For many smoothing methods, including ridge regression, we can write the model as $\boldsymbol{f} = \boldsymbol{h}\boldsymbol{\beta}$ where $\boldsymbol{h} = [\boldsymbol{h}_1, \dots, \boldsymbol{h}_n]^{\mathrm{T}}$ is an $n \times J$ design matrix with $\boldsymbol{h}_i$ a $1 \times J$ vector, and $\boldsymbol{\beta}$ is a $J \times 1$ vector of parameters. We prove the result (10.43) for a class of problems involving minimization of a sum of squares plus a quadratic penalty term:

$$\sum_{i=1}^{n} (y_i - \boldsymbol{h}_i\boldsymbol{\beta})^2 + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{\beta},$$

for a known matrix $\boldsymbol{D}$. Section 11.2.5 gives further examples of smoothers that fall within this class. Fitting the model to the $n-1$ points contained in $\boldsymbol{y}_{-i}$ involves minimization of

$$\sum_{j=1, j\neq i}^{n} [y_j - f_{-i}(\boldsymbol{x}_j)]^2 + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{\beta} = \sum_{j=1}^{n} \left[y_j^\star - f_{-i}(\boldsymbol{x}_j)\right]^2 + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{\beta} \tag{10.44}$$

where

$$y_j^\star = \begin{cases} y_j & \text{if } j \neq i \\ y_i - y_i + f_{-i}(\boldsymbol{x}_i) & \text{if } j = i. \end{cases}$$

Minimization of (10.44) yields

$$\widehat{\boldsymbol{f}} = \boldsymbol{S}^{(\lambda)}\boldsymbol{y}^\star = \boldsymbol{h}(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h} + \lambda\boldsymbol{D})^{-1}\boldsymbol{h}^{\mathrm{T}}\boldsymbol{y}^\star,$$

and $\widehat{f}_{-i}^{(\lambda)}(\boldsymbol{x}_i) = \boldsymbol{S}_i^{(\lambda)}\boldsymbol{y}^\star$, where $\boldsymbol{S}_i^{(\lambda)}$ is the ith row of $\boldsymbol{S}^{(\lambda)}$ and $\boldsymbol{y}^\star = [y_1^\star, \dots, y_n^\star]$. Now

$$\begin{aligned} \widehat{f}_{-i}^{(\lambda)}(\boldsymbol{x}_i) &= \boldsymbol{S}_i^{(\lambda)}\boldsymbol{y}^\star \\ &= \boldsymbol{S}_i^{(\lambda)}\boldsymbol{y} - S_{ii}^{(\lambda)}y_i + S_{ii}^{(\lambda)}\widehat{f}_{-i}^{(\lambda)}(\boldsymbol{x}_i) \\ &= \widehat{f}^{(\lambda)}(\boldsymbol{x}_i) - S_{ii}^{(\lambda)}y_i + S_{ii}^{(\lambda)}\widehat{f}_{-i}^{(\lambda)}(\boldsymbol{x}_i) \end{aligned}$$

so that

$$\widehat{f}_{-i}^{(\lambda)}(\boldsymbol{x}_i) = \frac{\widehat{f}^{(\lambda)}(\boldsymbol{x}_i) - S_{ii}^{(\lambda)}y_i}{1 - S_{ii}}$$

and

$$y_i - \widehat{f}_{-i}^{(\lambda)}(\boldsymbol{x}_i) = \frac{y_i(1 - S_{ii}^{(\lambda)}) - \widehat{f}^{(\lambda)}(\boldsymbol{x}_i) + S_{ii}^{(\lambda)}y_i}{1 - S_{ii}^{(\lambda)}} = \frac{y_i - \widehat{f}^{(\lambda)}(\boldsymbol{x}_i)}{1 - S_{ii}^{(\lambda)}}, \tag{10.45}$$

as required. To calculate the leave-one-out CV score, we therefore need only the residuals from the fit to the complete data and the diagonal elements of the smoother matrix. Note that the effect of $\left(1 - S_{ii}^{(\lambda)}\right)^2$ in the denominator of (10.43) is to inflate the residual at the i-th point, hence accounting for the underestimation of simply using the residual sum of squares. Formula (10.43) is true for all linear smoothers.

In practice, curves of the estimated prediction error against λ (the smoothing parameter) can be very flat, as shown for instance in Fig. 10.9. Therefore, as already noted, simply blindly using the value of λ that minimizes the cross-validation sum of squares is not a reliable strategy. In Hastie et al. (2009), it is recommended that λ be chosen such that the prediction error is no greater than one standard error above that with the lowest error. This approach results in a more parsimonious model being selected, though this recommendation is based on judgement and experience rather than theory.

10.6.3 Generalized Cross-Validation

So-called generalized cross-validation (GCV) provides an alternative to K-fold cross-validation. The GCV score is

$$\text{GCV}^{(\lambda)} = \frac{n}{\left[n - \text{tr}\left(\boldsymbol{S}^{(\lambda)}\right)\right]^2} \sum_{i=1}^{n} \left(y_i - \boldsymbol{S}_i^{(\lambda)} \boldsymbol{y}\right)^2 \tag{10.46}$$

for a linear smoother $\widehat{\boldsymbol{y}} = \boldsymbol{S}^{(\lambda)}\boldsymbol{y}$. An important early reference on the use of GCV is Craven and Wabha (1979). Recall that tr $\left(\boldsymbol{S}^{(\lambda)}\right)$ is the effective degrees of freedom of a linear smoother, (10.31), with larger values of λ corresponding to increased smoothing. Therefore, the denominator of (10.46) is the squared effective residual degrees of freedom and a measure of complexity: increasing λ decreases the effective number of parameters, that is, the complexity of the model, and this reduction produces lower variability. However, the numerator is the residual sum of squares and as such is a measure of squared bias with larger λ giving a poorer fit and increased bias. Consequently, we see that the GCV score is providing a trade-off between bias and variance. Unlike K-fold cross-validation, GCV does not require splitting of the data into cross-validation folds and repeatedly training and testing the model.

GCV may be justified/motivated in a number of different ways. On computational grounds, the GCV score is simpler to evaluate than the OCV score, since one only needs the trace of $\boldsymbol{S}^{(\lambda)}$ and not the diagonal elements $S_{ii}^{(\lambda)}$. Recall from Sect. 5.11.2 that in the context of a linear model, the *leverage* of y_i is defined as $S_{ii}^{(\lambda)}$, and so the OCV score can be highly influenced by a small number of data points (due to the presence of $1 - S_{ii}^{(\lambda)}$ in the denominator of (10.43)), which can be undesirable. Therefore, one interpretation of GCV is that it is simply a robust alternative to OCV with $1 - S_{ii}^{(\lambda)}$ replaced by $1 - \text{tr}(\boldsymbol{S}^{(\lambda)})/n$, which is clear if we rewrite (10.46) as

$$\begin{aligned}\text{GCV}^{(\lambda)} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\left(y_i - \boldsymbol{S}_i^{(\lambda)}\boldsymbol{y}\right)^2}{\left(1 - S_{ii}^{(\lambda)}\right)^2} \left(\frac{1 - S_{ii}^{(\lambda)}}{1 - \text{tr}\left(\boldsymbol{S}^{(\lambda)}\right)/n}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \left[y_i - \widehat{f}_{-i}^{(\lambda)}(\boldsymbol{x}_i)\right]^2 \left(\frac{1 - S_{ii}^{(\lambda)}}{1 - \text{tr}\left(\boldsymbol{S}^{(\lambda)}\right)/n}\right)^2 .\end{aligned}$$

This representation illustrates that those observations with large leverage are being down-weighted, as compared to OCV.

A final justification for using GCV, which was emphasized by Golub et al. (1979), is an invariance property. Namely, GCV is invariant to certain transformations of the data whereas OCV is not. Suppose we transform $\boldsymbol{y}$ and $\boldsymbol{x}$ to $\boldsymbol{Q}\boldsymbol{y}$ and $\boldsymbol{Q}\boldsymbol{x}$, respectively, where $\boldsymbol{Q}$ is any $n \times n$ orthogonal matrix (i.e., $\boldsymbol{Q}\boldsymbol{Q}^{\text{T}} = \boldsymbol{Q}^{\text{T}}\boldsymbol{Q} = \mathbf{I}_n$). For fixed λ, minimization with respect to $\boldsymbol{\beta}$ of

$$(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\text{T}}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\text{T}}\boldsymbol{\beta}$$

leads to inference that is identical to minimization of

$$(\boldsymbol{Q}\boldsymbol{y} - \boldsymbol{Q}\boldsymbol{x}\boldsymbol{\beta})^{\text{T}}(\boldsymbol{Q}\boldsymbol{y} - \boldsymbol{Q}\boldsymbol{x}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\text{T}}\boldsymbol{\beta}.$$

However, for fixed λ, the OCV scores are not identical, so that $\widehat{\lambda}$ obtained via minimization of the OCV will differ depending on whether we work with $\boldsymbol{y}$ or $\boldsymbol{Q}\boldsymbol{y}$.

If $\boldsymbol{S}^{(\lambda)}$ is the linear smoother for the original data, then

$$\boldsymbol{S}_Q^{(\lambda)} = \boldsymbol{Q}\boldsymbol{S}^{(\lambda)}\boldsymbol{Q}^{\text{T}}$$

is the linear smoother for the rotated data. Note that

$$\text{tr}\left(\boldsymbol{S}_Q^{(\lambda)}\right) = \text{tr}\left(\boldsymbol{Q}\boldsymbol{S}^{(\lambda)}\boldsymbol{Q}^{\text{T}}\right) = \text{tr}\left(\boldsymbol{S}^{(\lambda)}\boldsymbol{Q}^{\text{T}}\boldsymbol{Q}\right) = \text{tr}\left(\boldsymbol{S}^{(\lambda)}\right),$$

and GCV is invariant to the choice of $\boldsymbol{Q}$ (Golub et al. 1979). It can be shown (e.g., Wood 2006, Sect. 4.5.3) that GCV corresponds to the rotation of the data that results in each of the diagonal elements of $\boldsymbol{S}_Q^{(\lambda)}$ being equal. Since the expected prediction error is invariant to the rotation used, the GCV score shares with the OCV score the interpretation as an estimate of the expected prediction error.

Using the approximation $(1 - x)^{-2} \approx 1 + 2x$ we obtain

$$\begin{aligned}\text{GCV}^{(\lambda)} &\approx \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \widehat{f}^{(\lambda)}(\boldsymbol{x}_i)\right]^2 + \frac{2\text{tr}\left(\boldsymbol{S}^{(\lambda)}\right)}{n}\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \widehat{f}^{(\lambda)}(\boldsymbol{x}_i)\right]^2 \\ &= \frac{\text{RSS}^{(\lambda)}}{n} + \frac{2p^{(\lambda)}}{n}\widehat{\sigma}^2,\end{aligned}$$

which is proportional to Mallows C_P if we replace $\widehat{\sigma}^2_{\max}$ in (10.39) with $\widehat{\sigma}^2$, up to a constant not depending on λ.

10.6.4 AIC for General Models

The AIC was introduced in Sect. 4.8.2; here we provide a derivation as a generalization of Mallows C_P. Consider the prediction of new observations $Y_1^\star, \ldots, Y_n^\star$ with model

$$Y_i^\star \mid \boldsymbol{\beta} \sim_{ind} \mathrm{N}\left[f_i(\boldsymbol{\beta}), \sigma^2\right],$$

for $i = 1, \ldots, n$. Suppose we fit a model using data $\boldsymbol{Y}_n = [Y_1, \ldots, Y_n]$ and obtain the MLE $\widehat{\boldsymbol{\beta}}$. The expected value of the negative maximized log-likelihood evaluated at $\widehat{\boldsymbol{\beta}}$ is

$$-\mathrm{E}\left[l_n(\widehat{\boldsymbol{\beta}})\right] = \frac{n}{2}\log 2\pi + n \log\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^n \mathrm{E}\left\{\left[Y_i^\star - f_i(\widehat{\boldsymbol{\beta}})\right]^2\right\}.$$

Considering the last term only, we saw in Sect. 10.4.1 that

$$\sum_{i=1}^n \mathrm{E}\left\{\left[Y_i^\star - f_i(\widehat{\boldsymbol{\beta}})\right]^2\right\} = n\sigma^2 + \sum_{i=1}^n \mathrm{E}\left\{\left[f_i(\widehat{\boldsymbol{\beta}}) - f_i(\boldsymbol{\beta})\right]^2\right\}, \quad (10.47)$$

and Mallows C_P was derived as an approximation to the second term, with "good" models having a low C_p.

We now consider a general log-likelihood based on n observations $l_n(\boldsymbol{\beta})$, with our aim being to find a criterion to judge the "fits" of a collection of models, taking into account model complexity. The basis of AIC is to evaluate a model based on its ability to predict new data $Y_i^\star$, $i = 1, \ldots, n$. The prediction is based on the model $p(\boldsymbol{y}^\star \mid \widehat{\boldsymbol{\beta}})$ with $\widehat{\boldsymbol{\beta}}$ being the MLE based on an independent sample of size n, $\boldsymbol{Y}_n$.

The criterion that is used for discrimination, that is, to decide on whether the prediction is good, is the Kullback–Leibler distance (as discussed in Sect. 2.4.3) between the true model and the assumed model. The distance between the true (unknown) distribution $p_{\mathrm{T}}(\boldsymbol{y}^\star)$ and a model $p(\boldsymbol{y}^\star \mid \boldsymbol{\beta})$ is

$$\mathrm{KL}\left[p_{\mathrm{T}}(\boldsymbol{y}^\star), p(\boldsymbol{y}^\star \mid \boldsymbol{\beta})\right] = \int \log\left(\frac{p_{\mathrm{T}}(\boldsymbol{y}^\star)}{p(\boldsymbol{y}^\star \mid \boldsymbol{\beta})}\right) p_{\mathrm{T}}(\boldsymbol{y}^\star)\, d\boldsymbol{y}^\star \geq 0.$$

A good model with estimator $\widehat{\boldsymbol{\beta}}$ will produce a small value of

$$\mathrm{KL}\left[p_{\mathrm{T}}(\boldsymbol{y}^\star), p(\boldsymbol{y}^\star \mid \widehat{\boldsymbol{\beta}})\right]. \quad (10.48)$$

Unfortunately (10.48) cannot be directly used, since $p_{\mathrm{T}}(\boldsymbol{y}^\star)$ is unknown, but we show how it may be approximated, up to an additive constant.

Result: Let $\boldsymbol{Y}_n = [Y_1, \ldots, Y_n]$ be a random sample from $p_{\mathrm{T}}(y)$ and suppose a model $p(y \mid \boldsymbol{\beta})$ is fitted to these data and yields MLE $\widehat{\boldsymbol{\beta}}$, where $\boldsymbol{\beta}$ is a parameter vector of dimension p. For simplicity, we state and prove the result for independent and identically distributed data but the result is true in the nonidentically distributed case also. We wish to predict an independent sample, $Y_i^\star$, $i = 1, \ldots, n$, using $p(y^\star \mid \widehat{\boldsymbol{\beta}})$.

Two times the expected distance between the true distribution and the assumed distribution, evaluated at the estimator $\widehat{\boldsymbol{\beta}}$, is

$$\begin{aligned} D^\star &= 2 \times \mathrm{E}_{Y^\star}\left[\sum_{i=1}^{n} \log\left(\frac{p_{\mathrm{T}}(Y_i^\star)}{p(Y_i^\star \mid \widehat{\boldsymbol{\beta}})}\right)\right] \\ &= 2n \times \mathrm{KL}\left[p_{\mathrm{T}}(y^\star), p(y^\star \mid \widehat{\boldsymbol{\beta}})\right]. \end{aligned} \tag{10.49}$$

Then, we have the approximation

$$D^\star \approx 2n \times \mathrm{KL}\left[p_{\mathrm{T}}(y^\star), p(y^\star \mid \boldsymbol{\beta}_{\mathrm{T}})\right] + p, \tag{10.50}$$

where $\boldsymbol{\beta}_{\mathrm{T}}$ is the value of $\boldsymbol{\beta}$ that minimizes the Kullback–Leibler distance between $p_{\mathrm{T}}(\boldsymbol{y})$ and $p(\boldsymbol{y} \mid \boldsymbol{\beta})$ (for discussion, see Sect. 2.4.3). The difference between (10.49) and (10.50) therefore gives the increase in the discrepancy when $p(\boldsymbol{y}^\star \mid \boldsymbol{\beta}_{\mathrm{T}})$ is replaced by $p(\boldsymbol{y}^\star \mid \widehat{\boldsymbol{\beta}})$.

An estimate of $D^\star$ is

$$\widehat{D}^\star = -2 \times l_n(\widehat{\boldsymbol{\beta}}) + 2p + 2c_{\mathrm{T}}$$

where $c_{\mathrm{T}} = \int \log[p_{\mathrm{T}}(\boldsymbol{y}^\star)] p_{\mathrm{T}}(\boldsymbol{y}^\star)\, d\boldsymbol{y}^\star$ is a constant that is common to all models under comparison. Ignoring this constant gives Akaike's *An Information Criterion*[7] (AIC, Akaike 1973):

$$\mathrm{AIC} = -2 \times l_n(\widehat{\boldsymbol{\beta}}) + 2p.$$

Outline Derivation

The outline proof presented below is based on Davison (2003, Sect. 4.7). The distance measure $D^\star$ given in (10.49) is two times the expected difference between log-likelihoods:

$$D^\star = \mathrm{E}\left[2n \log p_{\mathrm{T}}(Y^\star) - 2n \log p(Y^\star \mid \widehat{\boldsymbol{\beta}})\right], \tag{10.51}$$

where the expectation is with respect to the true model $p_{\mathrm{T}}(\boldsymbol{y}^\star)$. We proceed by first approximating the second term via a Taylor series expansion about $\boldsymbol{\beta}_{\mathrm{T}}$. Let

$$\boldsymbol{S}_1(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log p(Y \mid \boldsymbol{\beta}), \quad \boldsymbol{I}_1(\boldsymbol{\beta}) = -\mathrm{E}\left[\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} \log p(Y \mid \boldsymbol{\beta})\right]$$

denote the score and information in a sample of size one. Then

$$\begin{aligned} 2n \log p(Y^\star \mid \widehat{\boldsymbol{\beta}}) \approx 2n \log p(Y^\star \mid \boldsymbol{\beta}_{\mathrm{T}}) &+ 2n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathrm{T}})^{\mathrm{T}} \boldsymbol{S}_1(\boldsymbol{\beta}_{\mathrm{T}}) \\ &- n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathrm{T}})^{\mathrm{T}} \boldsymbol{I}_1(\boldsymbol{\beta}_{\mathrm{T}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathrm{T}}). \end{aligned}$$

[7] Commonly AIC is referred to as *Akaike's Information Criterion*.

Note that $\text{E}\left[\boldsymbol{S}_1(\boldsymbol{\beta}_{\text{T}})\right] = \boldsymbol{0}$ and $n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\text{T}})^{\text{T}}\boldsymbol{I}_1(\boldsymbol{\beta}_{\text{T}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\text{T}})$ is asymptotically χ_p^2 (Sect. 2.9.4) so its expectation is p, the number of elements of $\boldsymbol{\beta}$. Hence, the second term in (10.51) may be approximated by

$$\text{E}\left[2n \log p(Y^\star \mid \widehat{\boldsymbol{\beta}})\right] \approx \text{E}\left[2n \log p(Y^\star \mid \boldsymbol{\beta}_{\text{T}})\right] - p. \tag{10.52}$$

Therefore,

$$\begin{aligned} D^\star &\approx 2n \times \text{E}\left[\log p_{\text{T}}(Y^\star) - \log p(Y^\star \mid \boldsymbol{\beta}_{\text{T}})\right] + p \\ &= 2n \int \log\left(\frac{p_{\text{T}}(y^\star)}{p(y^\star \mid \boldsymbol{\beta}_{\text{T}})}\right) p_{\text{T}}(y^\star)\, dy^\star + p \\ &= 2n \times \text{KL}\left[p_{\text{T}}(\boldsymbol{y}^\star), p(\boldsymbol{y}^\star \mid \boldsymbol{\beta}_{\text{T}})\right] + p \qquad (10.53) \end{aligned}$$

proving (10.50).

This expression for $D^\star$ is not usable because $p_{\text{T}}(\cdot)$ is unknown. An estimator of $\text{KL}\left[p_{\text{T}}(y^\star), p(y^\star \mid \boldsymbol{\beta}_{\text{T}})\right]$ can be based on $\text{E}\left[l_n(\widehat{\boldsymbol{\beta}})\right] = \text{E}\left[\log p(Y \mid \widehat{\boldsymbol{\beta}})\right]$, however. We write

$$\begin{aligned} -2 \times \text{E}\left[l_n(\widehat{\boldsymbol{\beta}})\right] &= 2 \times \text{E}\left[-l_n(\boldsymbol{\beta}_{\text{T}}) - \left\{ l_n(\widehat{\boldsymbol{\beta}}) - l_n(\boldsymbol{\beta}_{\text{T}}) \right\}\right] \\ &\approx 2n \times \text{E}\left[-\log p(Y^\star \mid \boldsymbol{\beta}_{\text{T}})\right] - p \\ &= 2n \times \text{E}\left[-\log p(Y^\star \mid \boldsymbol{\beta}_{\text{T}}) + \log p_{\text{T}}(Y^\star) - \log p_{\text{T}}(Y^\star)\right] - p \\ &= 2n \times \text{KL}\left[p_{\text{T}}(y^\star), p(\boldsymbol{y} \mid \boldsymbol{\beta}_{\text{T}})\right] - 2c_{\text{T}} - p \qquad (10.54) \end{aligned}$$

where

$$c_{\text{T}} = \int \log[p_{\text{T}}(\boldsymbol{y}^\star)]\, p_{\text{T}}(\boldsymbol{y}^\star)\, d\boldsymbol{y}^\star,$$

and we have used the asymptotic result that

$$2\left[l_n(\widehat{\boldsymbol{\beta}}) - l_n(\boldsymbol{\beta}_{\text{T}})\right] \to \chi_p^2, \tag{10.55}$$

as $n \to \infty$, see (2.55). It follows, by rearrangement of (10.54), that

$$2n \times \text{KL}\left[p_{\text{T}}(y^\star), p(y^\star \mid \boldsymbol{\beta}_{\text{T}})\right] \approx -2 \times \text{E}\left[l_n(\widehat{\boldsymbol{\beta}})\right] + p + 2c_T$$

which suggests an estimator of

$$2n \times \widehat{\text{KL}}\left[p_{\text{T}}(y^\star), p(y^\star \mid \boldsymbol{\beta}_{\text{T}})\right] = -2 \times l_n(\widehat{\boldsymbol{\beta}}) + p + 2c_{\text{T}}.$$

This estimate can be substituted into (10.53) to give the estimator

$$\begin{aligned} \widehat{D}^\star &= -2 \times l_n(\widehat{\boldsymbol{\beta}}) + 2p + 2c_{\text{T}} \\ &= \text{AIC} + 2c_{\text{T}} \end{aligned}$$

where AIC $= -2 \times l_n(\widehat{\boldsymbol{\beta}}) + 2p$. Since the term on the right is common to all models, the AIC may be used to compare models, with relatively good models producing a small value of the AIC. Some authors suggest retaining all models whose AIC is within 2 of the minimum (e.g. Ripley 2004). □

The above derivation is based on a number of assumptions (Ripley 2004) including the model under consideration being true. The accuracy of the approximations is also much greater if the models under comparison are nested.

In a GLM smoothing setting, the AIC may be minimized as a function of λ, with the degrees of freedom p being replaced by tr $\left(\boldsymbol{S}^{(\lambda)}\right)$. The AIC criteria in this case is

$$\text{AIC}^{(\lambda)} = -2l(\widehat{\boldsymbol{\beta}}) + 2 \times \text{tr}\left(\boldsymbol{S}^{(\lambda)}\right), \tag{10.56}$$

with the second term again measuring complexity.

An Aside

The derivation of AIC was carried out under the assumption of a correct model, which was required to obtain (10.52) and (10.55). If the model is wrong, then $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{\text{T}})$ is asymptotically normal with zero mean and variance $\boldsymbol{I}^{-1}\boldsymbol{K}\boldsymbol{I}^{\text{T}-1}$ where

$$\boldsymbol{K} = \boldsymbol{K}(\boldsymbol{\beta}_{\text{T}}) = \text{E}\left[\left(\frac{\partial}{\partial \boldsymbol{\beta}} \log p(Y \mid \boldsymbol{\beta}_{\text{T}})\right)\left(\frac{\partial}{\partial \boldsymbol{\beta}} \log p(Y \mid \boldsymbol{\beta}_{\text{T}})\right)^{\text{T}}\right],$$

see Sect. 2.4.3. Hence, using identity (B.4) from Appendix B, the expectation of $n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\text{T}})^{\text{T}}\boldsymbol{I}_1(\boldsymbol{\beta}_{\text{T}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\text{T}})$ is

$$\text{tr}\left[\boldsymbol{I}_1(\boldsymbol{\beta}_{\text{T}})\boldsymbol{I}_1(\boldsymbol{\beta}_{\text{T}})^{-1}\boldsymbol{K}(\boldsymbol{\beta}_{\text{T}})\boldsymbol{I}_1(\boldsymbol{\beta}_{\text{T}})^{-1}\right] = \text{tr}\left[\boldsymbol{K}(\boldsymbol{\beta}_{\text{T}})\boldsymbol{I}_1(\boldsymbol{\beta}_{\text{T}})^{-1}\right].$$

Similarly, under a wrong model, the likelihood ratio statistic $2\left[l_n(\widehat{\boldsymbol{\beta}}) - l_n(\boldsymbol{\beta}_{\text{T}})\right]$ has an asymptotic distribution proportional to χ^2_p but with mean tr $\left[\boldsymbol{K}(\boldsymbol{\beta}_{\text{T}})\boldsymbol{I}_1(\boldsymbol{\beta}_{\text{T}})^{-1}\right]$. This follows since, via a Taylor series approximation,

$$2\left[l_n(\widehat{\boldsymbol{\beta}}) - l_n(\boldsymbol{\beta}_{\text{T}})\right] \approx n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\text{T}})^{\text{T}}\boldsymbol{I}_1(\boldsymbol{\beta}_{\text{T}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\text{T}}).$$

Replacing p by tr $\left[\boldsymbol{K}(\boldsymbol{\beta}_{\text{T}})\boldsymbol{I}_1(\boldsymbol{\beta}_{\text{T}})^{-1}\right]$ in the above derivation gives the alternative *network information criterion* (NIC)

$$\text{NIC} = -2l(\widehat{\boldsymbol{\beta}}) + 2 \times \text{tr}\left[\boldsymbol{K}(\widehat{\boldsymbol{\beta}})\boldsymbol{I}_1(\widehat{\boldsymbol{\beta}})^{-1}\right],$$

as introduced by Stone (1977).

10.6.5 Cross-Validation for Generalized Linear Models

As discussed in Sect. 10.3.1, for general outcomes, a loss function for measuring the accuracy of a prediction is the negative log-likelihood. Hence, cross-validation can be extended to general data situations by replacing the sum of squares in (10.42) with a loss function to give

$$\text{CV}_k^{(\lambda)} = \frac{1}{n_k} \sum_{i \in J(k)} L\left[y_i, \widehat{f}_{-k}^{(\lambda)}(\boldsymbol{x}_i)\right].$$

In particular, the negative log-likelihood loss (10.12) produces

$$\text{CV}_k^{(\lambda)} = -\frac{2}{n_k} \sum_{i \in J(k)} \log p_{\widehat{f}_{-k}^{(\lambda)}}\left(y_i \mid \boldsymbol{x}_i\right),$$

where this notation emphasizes that the prediction at the point x_i is based upon the fitted value $\widehat{f}_{-k}^{(\lambda)}$. Similarly, a natural extension of (10.46) is the generalized cross-validation score based on the log-likelihood

$$\text{GCV}^{(\lambda)} = -\frac{2n}{\left[n - \text{tr}(\boldsymbol{S}^{(\lambda)})\right]^2} \sum_{i=1}^{n} \log p_{\widehat{f}_{-k}^{(\lambda)}}\left(y_i \mid \boldsymbol{x}_i\right).$$

Some authors (e.g., Ruppert et al. 2003, p. 220) replace the log-likelihood by the deviance, which adds a term that does not depend on λ.

Example: Prostate Cancer

We illustrate smoothing/tuning parameter choice and estimation of the prediction error using various approaches to modeling and a number of the methods described in Sect. 10.6 for smoothing parameter estimation. The modeling approaches we compare are fitting the full model using least squares, and picking the "best" subset of variables via an exhaustive search based on Mallows C_P, ridge regression, the lasso, and Bayesian model averaging (Sect. 3.6). We divide the prostate data into a training dataset of 67 randomly selected individuals and a test dataset of the remaining 30 individuals. Since the sample size is small, we repeat this splitting 500 times and then evaluate, for the different methods, the average error and its standard deviation over the train/test splits. An important point to emphasize is that we standardize the x variables in the training dataset and then apply the same standardization in the test dataset (and this procedure is repeated separately for each of the 500 splits).

Table 10.3 Average test errors over 500 train/test splits of the prostate cancer data, along with the standard deviation over these splits

	Null	Full	Best subset	Ridge	Lasso	BMA
Mean	1.30	0.59	0.76	0.59	0.60	0.59
SD	0.32	0.15	0.35	0.14	0.14	0.14

Table 10.3 gives summaries of the test error, calculated via (10.18), for the five approaches. We also report the error that results from fitting the null (intercept only) model. The latter is a baseline reference, and gives an error of 1.30. The estimate of error corresponding to the full model fitted with least squares is 0.59, a reduction of 71%. The exhaustive search over model space (i.e., the $2^8 = 256$ combinations of 8 variables), using Mallows C_P as the model selection criterion, was significantly worse giving an error of 0.76 with a large standard deviation. Table 10.4 shows the variability across train/test splits in the model chosen by the exhaustive search procedure. For example, 34.2% of models contained only the variables log(can vol), log(weight), and SVI. The seven most frequently occurring models account for 73.8% of the total, with the remainder being spread over 27 other combinations of variables. The table illustrates the discreteness of the exhaustive search procedure (as discussed in Sect. 4.9) and explains the poor prediction performance. Ridge regression and the lasso were applied to each train/test split with λ chosen via minimization of the OCV score. The entries in Table 10.3 show that, for these data, the shrinkage methods provide prediction errors which are comparable to, and not an improvement on, the full model. The reason for this is that in this example the ratio of the sample size to the number of parameters is relatively large, and so there is little penalty for including all parameters in the model.

Figure 10.8 illustrates the variability across train/test splits of the optimal effective degrees of freedom, chosen via minimization of (a) the OCV score and (b) Mallows C_P, for the ridge regression analyses. The two measures are then plotted against each other in (c) and show reasonable agreement. There is a reasonable amount of variability in the optimal degrees of freedom across simulations.

The final approach included in this experiment was Bayesian model averaging. In this example, the performance of BMA matches that of ridge regression and the lasso. BMA is superior to exhaustive search because covariates are not excluded entirely, but rather every model is assigned a posterior weight so that all covariates contribute to the fit. A number of successful approaches to prediction, including boosting, bagging, and random forests (Hastie et al. 2009), gain success from averaging over models, since different models can pick up different aspects of the data, and the variance is reduced by averaging. Bagging and random forests are described in Sects. 12.8.5 and 12.8.6, respectively.

We now provide more detail on the ridge regression, lasso, and Bayesian model averaging approaches. We first consider in greater detail the application of ridge regression. Figure 10.9 shows estimates of the test error, evaluated via different methods, as a function of the effective degrees of freedom, for a single train/test split. The minimizing values are indicated as vertical lines. The dotted line shows

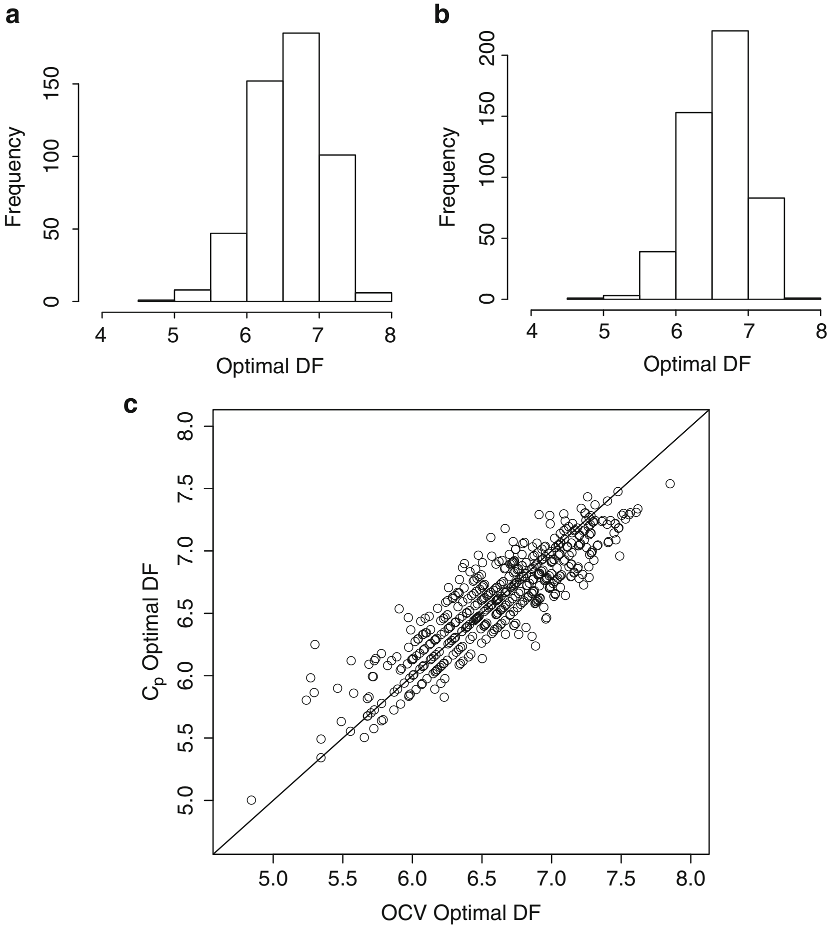


Fig. 10.8 Minimizing values of the effective degrees of freedom for ridge regression from 500 train/test splits of the prostate cancer data using: (**a**) OCV, (**b**) Mallows C_P as the minimizing criteria. Panel (**c**) plots the optimal degrees of freedom arising from each criteria against each other

the estimate as the AMSE plus the estimate of the error variance, (10.41). The minimizing value of AMSE (which is equivalent to minimizing Mallows C_P) is very similar to that obtained with the OCV criteria and is also virtually identical to that obtained from GCV. In all cases, the curves are flat close to the minimum, so one would not want to overinterpret specific numerical values. The effective degrees of freedom corresponding to the minimum OCV is 5.9, while under GCV and Mallows, the values are identical and equal to 5.7. The fivefold CV estimate is minimized for a slightly larger value than for OCV for this train/test split (effective degrees of freedom of 6.6); over all train/test splits, fivefold CV produced a comparable prediction error to OCV. Also included in the figure is the average residual sum of

Table 10.4 Percentage of models selected in an exhaustive best subset search, over 500 train/test splits of the prostate cancer data

Variables selected								
lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	Percentage
1	1	0	0	1	0	0	0	34.2
1	1	1	1	1	0	0	0	11.4
1	1	0	1	1	0	0	0	11.0
1	0	0	1	1	0	0	0	5.8
1	0	1	1	1	0	0	0	4.8
1	1	0	0	1	0	0	1	3.4
1	1	1	0	1	0	0	0	3.2

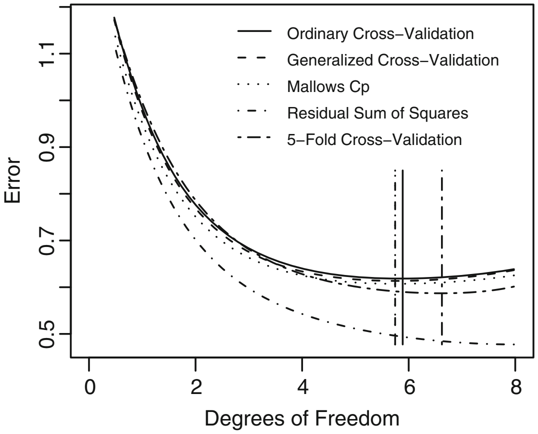


Fig. 10.9 Various estimates of error, as a function of the effective degrees of freedom, for ridge regression applied to the prostate cancer data. Minimizing values are shown as *vertical lines*. Also shown is the residual sum of squares (which has a minimum at 8 degrees of freedom)

squares, which is minimized at the most complex model (degrees of freedom equal to 8), as expected, and underestimates the predictive error, since the data are being used twice.

Turning now to the lasso, Figs. 10.10(a) and (b) show the OCV and GCV estimates of error versus the coefficient shrinkage factor, along with estimates of the standard error. As with ridge regression, the curves are relatively flat close to the minimum, indicating that we should not be wedded to the exact minimizing value of the smoothing parameter. For this train/test split, the minimizing value of the OCV function leads to three coefficients being set to zero.

Finally, for Bayesian model averaging, Fig. 10.11 provides a plot in which the horizontal axis orders the models in terms of decreasing posterior probability (going from left to right), with the variables indicated on the vertical axis. Black rectangles denote inclusion of that variable and gray, no inclusion. The posterior model percentages for the top five models are 23%, 17%, 8%, 7%, and 6%.

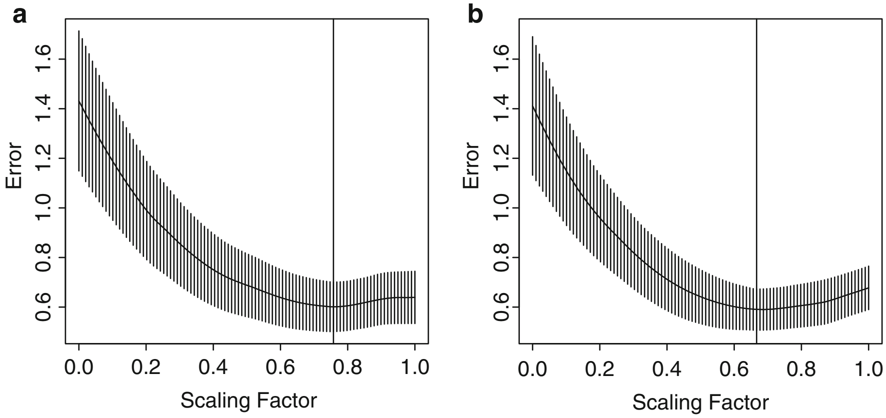
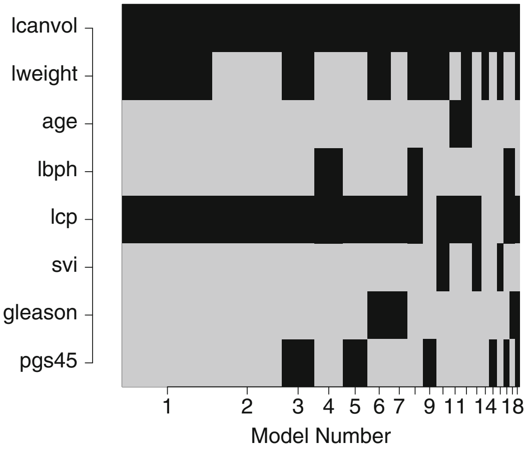


Fig. 10.10 (**a**) OCV and (**b**) fivefold CV estimates of error for the lasso, as a function of the scaling factor, $\sum_{j=1}^{k} |\widehat{\beta}_j| / \sum_{j=1}^{k} |\widehat{\beta}_j^{\text{LS}}|$, for the prostate cancer data. The minimizing value of the CV estimates of error is shown as a *solid vertical line*. Also shown are approximate standard error bands evaluated as if the CV estimates were independent (as discussed in Sect. 10.6.2)

Fig. 10.11 From left to right this plot shows, for a particular split of the prostate cancer data, the models with the highest posterior probability, as evaluated via Bayesian model averaging



10.7 Concluding Comments

Whether parametric or nonparametric models are used, the bias-variance trade-off is a key consideration. In nonparametric modeling there are explicit smoothing parameters that determine this trade-off. We saw this with both ridge regression and the lasso, and this issue will return repeatedly in Chaps. 11 and 12. The choice of smoothing parameter is, therefore, crucial and a variety of approaches for selection, including cross-validation and the minimization of Mallows C_P have been described. Additional methods will be described in Chap. 11, but no single approach will work in all situations, and often subjective judgement is required. Härdle et al.

(1988) have shown that smoothing parameter methods such as Mallows C_P and GCV converge slowly to the optimum as the sample size increases. A number of simulation studies have been carried out and back up the above comments, see, for example, Ruppert et al. (2003, Sect. 5.4) and references therein.

10.8 Bibliographic Notes

There are many excellent texts on nonparametric regression, including Green and Silverman (1994), Simonoff (1997), Ruppert et al. (2003), Wood (2006), and, more recently and with a large range of topics, Hastie et al. (2009). Gneiting and Raftery (2007) provide an excellent review of scoring rules, which are closely related to the loss functions considered in Sect. 10.3. An important early reference on ridge regression is Hoerl and Kennard (1970). Since its introduction in Tibshirani (1996), the lasso has been the subject of much interest, see Tibshirani (2011) and the ensuing discussion for a summary. There is a considerable literature on the theoretical aspects of the lasso, for example, examining its properties with respect to prediction loss and model selection, see Meinshausen and Yu (2009) and references therein.

10.9 Exercises

10.1 For the LIDAR data described in Sect. 10.2.1 fit polynomials of increasing degree as a function of range and comment on the fit to the data. These data are available in the R package `SemiPar` and are named `lidar`. What degree of polynomial is required to obtain an adequate fit to these data?
[Hint: One method of assessing the latter is to examine residuals.]

10.2 The BPD data described in Sect. 7.2.3 are available on the book website. Fit linear and quadratic logistic regression models to these data and interpret the parameters.

10.3 Carry out backwards elimination for the prostate cancer data, which are available in the R package `lasso2` and are named `Prostate`. Comment on the standard errors of the estimates in the final model that you arrive at, as compared to the corresponding estimates in the full model.

10.4 With reference to Sect. 10.3.1:

a. Show that minimization of expected quadratic loss, $\mathrm{E}_{\boldsymbol{X},Y}\left\{[Y - f(\boldsymbol{X})]^2\right\}$ leads to $\widehat{f}(\boldsymbol{x}) = \mathrm{E}[Y \mid \boldsymbol{x}]$.

b. Show that minimization of expected absolute value loss, $\mathrm{E}_{\boldsymbol{X},Y}[\ |Y - f(\boldsymbol{X})|\ ]$ leads to $\widehat{f}(\boldsymbol{x}) = \mathrm{median}(Y \mid \boldsymbol{x})$.

c. Consider the bilinear loss function

$$L[y,f(\boldsymbol{x})] = \begin{cases} a\,[y-f(\boldsymbol{x})] & \text{if } f(\boldsymbol{x}) \le y \\ b\,[f(\boldsymbol{x})-y] & \text{if } f(\boldsymbol{x}) \ge y. \end{cases}$$

Deduce that this leads to the optimal $f(\boldsymbol{x})$ being the $100 \times a/(a+b)\%$ point of the distribution function of Y.

10.5 a. Show that the expected value of scaled quadratic loss

$$\mathrm{E}_{Y \mid \boldsymbol{x}} \left\{ \frac{[Y-f(\boldsymbol{x})]^2}{Y^2} \right\}$$

is minimized by

$$\widehat{f}(\boldsymbol{x}) = \frac{\mathrm{E}[Y^{-1} \mid \boldsymbol{x}]}{\mathrm{E}[Y^{-2} \mid \boldsymbol{x}]}.$$

b. Suppose $Y \mid \mu(\boldsymbol{x}), \alpha \sim \mathrm{Ga}\left\{\alpha^{-1}, [\mu(\boldsymbol{x})\alpha]^{-1}\right\}$ and that prediction of Y using $f(\boldsymbol{x})$ is required, under scaled quadratic loss. Show that $\widehat{f}(\boldsymbol{x}) = \mathrm{E}[Y^{-1} \mid \boldsymbol{x}] = (1-2\alpha)\mu(\boldsymbol{x})$.
[Hint: If $Y \mid a,b \sim \mathrm{Ga}(a,b)$, then $Y^{-1} \mid a,b \sim \mathrm{InvGa}(a,b)$.]

10.6 From Sect. 10.5.1 show, using a Lagrange multiplier argument, that minimizing the penalized sum of squares:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{k} \beta_j^2,$$

is equivalent to minimization of

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j \right)^2$$

subject to

$$\sum_{j=1}^{k} \beta_j^2 \le s,$$

for some s.

10.7 Prove the alternative formulas (10.25)–(10.27) for ridge regression.

10.8 Show, using (10.45), that

$$\mid y_i - \widehat{f}_{-i}^{(\lambda)}(\boldsymbol{x}_i) \mid \ \ge \ \mid y_i - \widehat{f}^{(\lambda)}(\boldsymbol{x}_i) \mid.$$

Interpret this result.

10.9 Cross-validation can fail completely for some problems, as will now be illustrated.

(a) Suppose we smooth a response y_i, by minimizing, with respect to μ_i, $i = 1, \ldots, n$, the ridge regression sum of squares

$$\sum_{i=1}^{n}(y_i - \mu_i)^2 + \lambda \sum_{i=1}^{n} \mu_i^2,$$

where λ is the smoothing parameter. Show that for this problem, the OCV and GCV scores are identical and independent of λ.

(b) By considering the basic principle of OCV, explain what causes the failure of the previous part.

(c) Given the explanation of the failure of cross-validation for the ridge regression problem in part (a), it might be expected that the following modified approach will work better. Suppose a covariate x_i is observed for each y_i (and for convenience, assume $x_i < x_{i+1}$ for all i). Define $\mu(x)$ to be the piecewise linear function with $n - 1$ linear segments between x_i and x_{i-1} for $i = 2, \ldots, n$. In this case μ_i could be estimated by minimizing the following penalized least squares objective:

$$\sum_{i=1}^{n}(y_i - \mu_i)^2 + \lambda \int \mu(x)^2 dx,$$

with respect to μ_i, $i = 1, \ldots, n$.

Now consider three equally spaced points x_1, x_2, x_3 with corresponding μ values μ_1, μ_2, μ_3. Suppose that $\mu_1 = \mu_3 = \mu^\star$, but that μ_2 can be freely chosen. Show that in order to minimize $\int_{x_1}^{x_3} \mu(x)^2 dx$, μ_2 should be set to $-\mu^\star/2$. What does this imply about trying to choose λ by cross-validation?

[Hint: think about what the penalty will do to μ_i if we “leave out” y_i.]

(d) Would the penalty

$$\int \mu'(x)^2 \, dx$$

suffer from the same problem as the penalty used in part (c)?

(e) Would you expect to encounter these sorts of problems with penalized regression smoothers? Explain your answer.

10.10 In this question data in the R package faraway that are named meatspec will be analyzed. Theses data concern the fat content, which is the response, measured in 215 samples of finely chopped meat, along with 100 covariates measuring the absorption at 100 wavelengths. Perform ridge regression on these data using OCV and GCV to choose the smoothing parameter. You should include a plot of how the estimates change as a function of the smoothing parameter and a plot displaying the cross-validation scores as a function of the smoothing parameter.

10.11 For the prostate cancer data considered throughout this chapter, reproduce the summaries in Table 10.3, coding up “by hand” the cross-validation procedures.